Published in final edited form as:
Nature. 2020 April ; 580(7802): 245–251. doi:10.1038/s41586-020-2140-0.

Integrating genomic features for noninvasive early lung cancer detection

Jacob J. Chabon[#1,2], **Emily G. Hamilton**[#3], **David M. Kurtz**[#4,5,6], **Mohammad S. Esfahani**[#1,4], **Everett J. Moding**[1,7], **Henning Stehr**[8], **Joseph Schroers-Martin**[4], **Barzin Y. Nabet**[1], **Binbin Chen**[4,9], **Aadel A. Chaudhuri**[10,11,12], **Chih Long Liu**[4], **Angela B. Hui**[1,7], **Michael C. Jin**[4], **Tej D. Azad**[4], **Diego Almanza**[3], **Young-Jun Jeon**[1], **Monica C. Nesselbush**[3], **Lyron Co Ting Keh**[1], **Rene F. Bonilla**[7], **Christopher H. Yoo**[7], **Ryan B. Ko**[7], **Emily L. Chen**[7], **David J. Merriott**[7], **Pierre P. Massion**[13,14], **Aaron S. Mansfield**[15], **Jin Jen**[16], **Hong Z. Ren**[16], **Steven H. Lin**[17], **Christina L. Costantino**[18,19], **Risa Burr**[18,20], **Robert Tibshirani**[21,22], **Sanjiv S. Gambhir**[6,24], **Gerald J. Berry**[8], **Kristin C. Jensen**[8,23], **Robert B. West**[8], **Joel W. Neal**[4], **Heather A. Wakelee**[4], **Billy W. Loo Jr**[7], **Christian A. Kunder**[8], **Ann N. Leung**[24], **Natalie S. Lui**[25], **Mark F. Berry**[25], **Joseph B. Shrager**[23,25], **Viswam S. Nair**[24,26,27], **Daniel A. Haber**[18,20,28], **Lecia V. Sequist**[18,28], **Ash A. Alizadeh**[1,2,4,5,#], **Maximilian Diehn**[1,2,7,#]

[1]Stanford Cancer Institute, Stanford University, Stanford, CA

[2]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA

[3]Program in Cancer Biology, Stanford University, Stanford, CA

[4]Division of Oncology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, CA

[5]Division of Hematology, Department of Medicine, Stanford University, Stanford, CA

[6]Department of Bioengineering, Stanford University, Stanford, CA

[7]Department of Radiation Oncology, Stanford University, Stanford, CA

[8]Department of Pathology, Stanford University, Stanford, CA

Corresponding Authors: Maximilian Diehn, MD/PhD, Department of Radiation Oncology, Cancer Institute, and Institute for Stem Cell Biology & Regenerative Medicine, 265 Campus Drive, Room G2120A, Stanford, CA 94305. diehn@stanford.edu; Ash A. Alizadeh, MD/PhD, Divisions of Oncology & Hematology, Cancer Institute, Institute for Stem Cell Biology, 265 Campus Drive, Room G2120B, Stanford, CA 94305, arasha@stanford.edu. [#]Co-corresponding authors.

Author Information

These authors contributed equally: Jacob J. Chabon, Emily G. Hamilton, David M. Kurtz, and Mohammad S. Esfahani. Maximilian Diehn and Ash A. Alizadeh jointly supervised this work. Requests for additional materials related to this work should be directed to Maximilian Diehn.

Author contributions

J.J.C, E.G.H, D.M.K., M.S.E, A.A.A. and M.D. developed the concept, designed the experiments, and analyzed the data. J.J.C, E.G.H, D.M.K., M.S.E, A.A.A. and M.D. wrote the manuscript. J.J.C. and D.M.K. developed the *in silico* simulation of the CAPP-Seq workflow and the FLEX adapters with input from M.D. and A.A.A. J.J.C and M.S.E developed the machine learning module of the Lung-CLiP model with input from E.G.H and D.M.K. J.J.C. performed molecular biology experiments related to improving the CAPP-Seq workflow. J.J.C. and E.G.H. performed molecular biology related to profiling clinical specimens with assistance from E.J.M., B.Y.N, A.A.C., A.B.H., T.D.A., Y.J.J., M.C.N. and D.A. Bioinformatics analyses were performed by J.J.C., E.G.H, D.M.K., M.S.E., H.S., J.S.M., B.C., C.L.L., M.C.J., M.C.N. and L.C.T. Patient specimens were provided by P.P.M., A.S.M., J.J., S.H.L, C.L.C., R.B., J.W.N., H.A.W., B.W.L., N.S.L., M.F.B, J.B.S., S.S.G., V.S.N., D.A.H. and L.V.S., and M.D. A.N.L. performed radiologic analyses. R.F.B, C.H.Y., R.B.K., E.L.C, D.J.M., P.P.M, H.Z.R., A.S.M., C.L.C., R.B., G.J.B., K.C.J., R.B.W., C.A.K. and V.S.N. organized patient enrollment, sample collection, and clinical data curation. All authors reviewed the manuscript. A.A.A. and M.D. contributed equally as senior authors.

[9]Department of Genetics, Stanford University, Stanford, CA

[10]Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO

[11]Department of Genetics, Washington University School of Medicine, St. Louis, MO

[12]Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO

[13]Division of Allergy, Pulmonary and Critical Care Medicine, Vanderbilt University Medical Center, Nashville, TN

[14]Veterans Affairs, Tennessee Valley Healthcare System, Nashville, TN

[15]Department of Oncology, Division of Medical Oncology, Mayo Clinic, Rochester, MN

[16]Division of Experimental Pathology, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN

[17]Department of Radiation Oncology, University of Texas MD Anderson Cancer Center, Houston, TX

[18]Massachusetts General Hospital Cancer Center, Harvard Medical School, Boston, MA

[19]Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA

[20]Howard Hughes Medical Institute, Chevy Chase, MD

[21]Department Statistics, Stanford University, Stanford, CA

[22]Department of Biomedical Data Science, Stanford University, Stanford, CA, University, Stanford, California, USA.

[23]VA Palo Alto Healthcare System, Palo Alto, CA

[24]Department of Radiology, Stanford University, Stanford, CA

[25]Division of Thoracic Surgery, Department of Cardiothoracic Surgery, Stanford

[26]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA

[27]Division of Pulmonary, Critical Care and Sleep Medicine, University of Washington, Seattle, WA

[28]Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA

[#] These authors contributed equally to this work.

Summary

Radiologic screening of high-risk adults reduces lung cancer-related mortality[1–3]. Despite this, < 5% of eligible individuals undergo such screening in the U.S.[4,5], and the availability of blood-based tests could increase adoption. Here, we introduce enhancements to Cancer Personalized Profiling by deep Sequencing (CAPP-Seq)[6] circulating tumor DNA (ctDNA) analysis that facilitate screening applications. We show that although ctDNA levels are very low in early-stage lung cancers, ctDNA is present pre-treatment in most patients and is strongly prognostic. We also demonstrate that the majority of somatic mutations in cell-free DNA (cfDNA) of lung cancer patients and risk-matched controls reflect clonal hematopoiesis (CH) and are non-recurrent.

Compared to tumor-derived mutations, CH mutations occur on longer cfDNA fragments and lack mutational signatures associated with tobacco smoking. Integrating these findings with other molecular features, we develop and prospectively validate a machine learning-based Lung Cancer Likelihood in Plasma (Lung-CLiP) method that robustly discriminates early-stage lung cancer patients from risk-matched controls. Our approach achieves similar performance as tumor-informed ctDNA detection and allows for tuning of specificity to facilitate distinct clinical applications. Our findings establish the potential of cfDNA for lung cancer screening and highlight the importance of risk-matching cases and controls in cfDNA-based screening studies.

Although annual radiologic screening via low-dose computed tomography (LDCT) is recommended for high-risk populations in the United States[3], implementation has been complicated by a high false discovery rate (~90%)[1] and low compliance[4,5]. Therefore, there is an unmet need for new methods for early detection of lung cancers. Analysis of ctDNA is a promising approach that could facilitate blood-based screening.

Improving detection of rare variants

Prior studies have shown that most patients with stage I lung cancer harbor ctDNA levels below 0.1%[7,8]. We therefore began by enhancing our previously described CAPP-Seq method[6] for early detection applications (Fig. 1a, Supplementary Methods, Supplemental Note). Specifically, using *in silico* simulations of molecular losses at various workflow steps, we optimized our protocol to improve recovery of unique cfDNA molecules and the fraction of cfDNA duplexes for which both strands were sequenced (Extended Data Fig. 1). Separately, we improved the error profile of our sequencing data through chemical inhibition of G-oxidation that occurs during hybrid capture enrichment (Extended Data Fig. 2). Lastly, we developed a custom duplex adapter schema for library preparation with several advantages compared to our previously described tandem adapters[6] (Extended Data Fig. 3).

Tumor-informed ctDNA detection

As a step towards developing a noninvasive method for non-small cell lung cancer (NSCLC) screening, we aimed to determine ctDNA detection rates in early-stage NSCLC patients using a tumor-informed approach. We genotyped tumor tissue, pre-treatment plasma cfDNA, and leukocyte DNA from 85 patients with stage I-III NSCLC by targeted deep sequencing of 255 genes recurrently mutated in NSCLC using a 355 kb CAPP-Seq panel (Extended Data Fig. 4). Using this panel, which does not require patient-specific customization (i.e. ‘population-based’ approach), we found a median of 4 mutations per patient in tumor specimens and detected ctDNA in 49% (38/85) of patients. We found that sensitivity of detection improved as the number of monitored tumor mutations increased (Extended Data Fig. 5a–b). To empirically test whether tracking more mutations improves sensitivity, we designed customized capture panels based on tumor exome sequencing data for 17 patients in whom ctDNA was not detected using the population-based panel. Using these customized panels, we detected ctDNA in 10/17 (59%) patients at a median VAF of 0.002% and at levels as low as 2.9 in 10^6 molecules (Extended Data Fig. 5c).

Combining the results of population-based (n=68) and customized (n=17) tumor-informed strategies, we detected ctDNA in 42%, 67%, and 88% of patients with stage I, II, and III disease, respectively (Fig. 1b). The patient-specific analytical limit of detection (LOD) was inferior in patients without detectable ctDNA (Extended Data Fig. 5d), suggesting that detection might improve by increasing the number of mutations monitored or the unique molecular depth. Indeed, when considering only patients for whom a LOD of < 0.01% was achieved, sensitivity increased to 63%, 82%, and 100% for stage I, II, and III tumors, respectively. Strikingly, we found that 50%, 38% and 7% of stage I, II, and III patients had ctDNA levels below 0.01%, respectively (Fig. 1c). Thus, the majority of localized NSCLCs shed ctDNA, but levels in many cases are lower than previously recognized[7,8].

We characterized properties of ctDNA molecules that could inform tumor-naïve screening. Consistent with prior reports[8–10], clonal tumor mutations were more frequently detected in plasma and observed at higher VAFs than their subclonal counterparts (Extended Data Fig. 5e), and cfDNA molecules harboring mutations present in matched tumor samples (i.e. ctDNA) were shorter than non-mutant molecules (Fig. 1d). Mutant cfDNA molecules were enriched among sub-mononucleosomal and sub-disomal fragments (Fig. 1e). When only considering molecules in these size windows, we observed a 2.17-fold median enrichment in the VAFs of tumor-derived mutations (Extended Data Fig. 5f–g), suggesting that *in silico* size selection for these regions might be useful. However, size selection disproportionately favored variants with higher pre-enrichment VAFs (Extended Data Fig. 5h–i), and although size selection improved sensitivity when using customized panels, sensitivity degraded when using the population-based panel (Extended Data Fig. 5j). This suggests that considering the extent to which a mutation is enriched in these regions may have advantages over only considering molecules in the ctDNA-enriched size windows.

Clinical correlates of ctDNA detection

We next evaluated clinical and pathological correlates of ctDNA levels. We found ctDNA level to be associated with stage (Fig. 1f), metabolic tumor volume (MTV) (Fig. 1g, Extended Data Fig. 6a–b), and tumor histology (Fig. 1h). Interestingly, each of these parameters were independently associated with ctDNA level in multiple variable analysis (Extended Data Fig. 6c), suggesting that ctDNA levels reflect multiple biological factors.

Lung adenocarcinomas exist on a spectrum from pre-invasive to frankly invasive epithelial proliferations, associated with differences in radiologic appearance ranging from pure ground-glass opacities (GGOs) to solid lesions. Since GGO-predominant lung cancers are slow growing and often indolent[11], we hypothesized that they shed less ctDNA than solid lesions. We detected ctDNA less frequently and at lower levels in patients with a substantial ground-glass component (≥ 25% GGO, Extended Data Fig. 6d). Separately, ctDNA was more frequently detectable in patients whose tumors displayed radiologic evidence of necrosis (Extended Data Fig. 6e). Thus, imaging characteristics of NSCLCs are associated with ctDNA shedding and may help identify patients most appropriate for ctDNA analysis.

Given that prior studies have found that residual ctDNA following treatment of localized NSCLC portends a high risk of recurrence[7,8,12], we next tested the association of pre-treatment ctDNA levels with clinical outcomes. Patients with higher-than-median ctDNA levels had inferior freedom from recurrence (Fig. 1i) and recurrence-free survival (Extended Data Fig. 7a). Pre-treatment ctDNA level was similarly prognostic when only considering patients with stage I disease (Fig. 1j, Extended Data Fig. 7b). Importantly, in multivariable analysis including both MTV and stage, only ctDNA was significantly associated with outcome (Fig. 1k). Since distant metastasis drives cancer-associated mortality after treatment of localized NSCLC, we examined the association of pre-treatment ctDNA levels with future metastasis. Strikingly, high pre-treatment ctDNA was also associated with inferior freedom from distant metastasis (Extended Data Fig. 7c–e). Thus, pre-treatment ctDNA level is a previously unrecognized prognostic factor in localized NSCLC that appears to enrich for patients harboring micro-metastatic disease (Extended Data Fig. 7f).

Sources of cfDNA somatic variants

Clonal hematopoiesis (CH) is an aging-related phenomenon wherein non-malignant hematopoietic stem/progenitor cells acquire somatic alterations that can confer a selective advantage[13]. Since hematopoietic cells are the primary source of cfDNA[14] and contribute somatic variants to the cfDNA pool[15,16], we sought to identify approaches for distinguishing CH-derived mutations from their tumor-derived counterparts.

We began by examining whether variants found in cfDNA were also detected in matched white-blood cells (WBCs) in early-stage NSCLC patients (n=104) and non-cancer control subjects (n=98). We considered two separate control groups: 56 age-, sex- and smoking status-matched adults undergoing annual radiologic screening for lung cancer (“risk-matched controls”), and 42 un-matched adult blood donors (“low-risk controls,”). We observed more total cfDNA mutations and mutations that were absent in matched leukocytes (i.e. “WBC-”) in NSCLC patients than both control groups (Fig. 2a). However, both NSCLC patients and risk-matched controls harbored more cfDNA mutations and CH variants (i.e. “WBC+”) than low-risk controls, highlighting the importance of risk-matching cases and controls in cfDNA-based early detection studies.

We found that 94.8% of WBC+ cfDNA mutations were private to individual subjects (Fig. 2b) and 48% of WBC+ cfDNA mutations in controls affected genes not canonically associated with CH (Fig. 2c). Importantly, the majority of cfDNA variants in both NSCLC patients (58%) and controls (90%) were attributable to CH and in 76% of patients and 91% of controls the mutation with the highest VAF was also present in matched WBCs (Fig. 2d). The VAFs of mutations observed in both compartments were significantly correlated (Fig. 2e) and 81% of WBC+ cfDNA variants had VAFs below 1% in leukocytes (Extended Data Fig. 8a). These findings highlight the importance of sequencing matched leukocyte DNA and cfDNA to equivalent depths to determine whether cfDNA mutations are CH-derived.

In individuals without a hematologic neoplasm, WBC mutations in canonical CH genes with a VAF ≥ 2% are commonly referred to as clonal hematopoiesis of indeterminate potential (CHIP)[13]. We observed one or more such mutations in 13.5% of NSCLC patients, 7.1% of risk-matched controls, and none of the low-risk controls. As expected, variants in WBCs occurring at ≥ 2% VAF more frequently affected canonical CH genes than variants occurring at < 2% (Extended Data Fig. 8b).

Since CH is known to increase with age[13], we next examined whether the number of WBC+ cfDNA mutations was associated with age. We found that the number of WBC+ mutations, but not WBC- mutations, was significantly correlated with age (Fig. 2f). Consistent with the concept that these mutations constitute CH events, the genes most frequently containing WBC+ cfDNA mutations were canonical CH genes, including *DNMT3A*, *TET2*, *TP53*, *PPM1D* and *SF3B1* (Extended Data Fig. 8c–d).

To examine whether presence of WBC+ cfDNA mutations changed over time, we considered the subset of our cohort with plasma from two time points. Among WBC+ cfDNA mutations detected at the first time point, 74% (42/57) were also detected at the second time point and had highly correlated VAFs (Extended Data Fig. 8e). Additionally, considering all WBC+ cfDNA mutations, canonical CH genes harbored higher rates of nonsynonymous mutations than synonymous variants (Extended Data Fig. 8f), consistent with these mutations being under positive selection.

We next compared the mutational signatures of WBC+ and WBC- cfDNA mutations to each other and to mutation datasets from the CH and lung cancer literature[17–19]. We found that WBC+ cfDNA mutations were dominated by the aging-associated mutational signature (Signature 1) in both NSCLC patients and controls (Fig. 2g). Interestingly, Signature 4, which is associated with tobacco smoking and is the predominant mutational signature of NSCLC tumor genomes[20], was observed in WBC- but not WBC+ cfDNA mutations in NSCLC patients and was not observed in either compartment among controls with or without a history of smoking. This suggests that the base substitution spectrum of cfDNA variants might be useful for distinguishing carcinoma-derived from CH-derived mutations.

TP53 is the most frequently mutated gene in human cancers[21]; however, mutations in *TP53* are also frequently seen in CH[17]. Discrimination between carcinoma-derived and CH-derived *TP53* mutations is therefore an important consideration for cfDNA-based cancer screening approaches. Notably, many *TP53* variants found in cfDNA were also detectable in WBCs in both NSCLC patients (40.6%) and controls (100%; Extended Data Fig. 8c). Although the distribution of WBC+ and WBC- cfDNA mutations was similar across the p53 protein (Extended Data Fig. 8g), WBC- *TP53* mutations displayed stronger evidence of the smoking mutational signature than their WBC+ counterparts (Fig. 2h).

We also studied the fragment size distribution of cfDNA molecules harboring variants present in matched WBCs or in matched tumor biopsies. We found that cfDNA molecules harboring WBC+ cfDNA mutations displayed a nearly identical size distribution as non-mutant molecules (Fig. 2i). In contrast, cfDNA molecules harboring mutations present in matched tumor specimens were significantly shorter than non-mutant molecules. Accordingly, *in silico* size selection for the fragment sizes found to be ctDNA-enriched in our tumor-informed analysis (Fig. 1e) did not increase the VAFs of CH mutations in NSCLC patients or controls (Extended Data Fig. 8h). However, the VAFs of WBC- mutations in NSCLC patients, but not in controls, significantly enriched with size selection. This suggests that cfDNA fragment size may also be useful for distinguishing carcinoma-derived from CH-derived mutations.

Estimating cancer likelihood in plasma

Having identified properties that distinguish tumor-derived and CH-derived cfDNA fragments, we developed a method to measure Lung Cancer Likelihood in Plasma (Lung-CLiP). This approach involves targeted sequencing of plasma cfDNA and matched leukocyte DNA and integrates single nucleotide variants (SNVs) and genome-wide copy number analysis with machine learning models. We trained Lung-CLiP using samples from a discovery cohort of 104 early-stage NSCLC patients and 56 risk-matched controls undergoing annual radiologic screening for lung cancer at 4 cancer centers. To develop Lung-CLiP we employed a multi-tiered machine learning approach in which we first trained a model to estimate the probability a cfDNA mutation is tumor-derived. This ‘SNV model’ leverages biological and technical features specific to each variant such as background frequencies, cfDNA fragment size, the gene affected, and CH-likelihood (Extended Data Fig. 9a). Next, we utilize both the on- and off-target sequencing reads from CAPP-Seq to identify genome-wide copy number alterations. The results of the SNV model and the genome-wide copy number calls are then integrated within a final patient-level classifier that estimates the likelihood a blood sample contains lung cancer-derived cfDNA (i.e. “Lung-CLiP score,” Fig. 3a).

Receiver-operator characteristic curve shapes revealed that Lung-CLiP sensitivity can be tuned to desirable specificities depending on the target clinical application (Extended Data Fig. 9b). For example, as a standalone screening test, high specificity would be desirable to minimize false positives. At 98% specificity, we observed sensitivities of 41% in stage I, 54% in stage II, and 67% in stage III patients (Fig. 3b). Alternatively, a lower specificity may be acceptable if Lung-CLiP were applied to the ~95% of at-risk individuals who are not currently undergoing LDCT screening due to access limitations or other hurdles[4,5]. In this context, a lower specificity would be reasonable since the reflex test for a positive Lung-CLiP test would be LDCT. For example, at 80% specificity we observed sensitivities of 63% in stage I, 69% in stage II, and 75% in stage III patients (Fig. 3c).

To confirm the biological plausibility of Lung-CLiP scores, we compared them to tumor-informed ctDNA levels and clinicopathological features (Fig. 3d). Lung-CLiP achieved statistically similar stage-matched sensitivities at 98% specificity as tumor-informed ctDNA analysis (Fig. 3e) and Lung-CLiP scores were correlated with tumor-informed ctDNA levels (Fig. 3f). As expected, tumors from NSCLC patients classified as positive by Lung-CLiP were larger than those classified as negative (Fig. 3g), and patients with non-adenocarcinoma histology were more frequently detected (Fig. 3h). Taken together, these data suggest that Lung-CLiP scores capture biologically meaningful factors related to overall ctDNA burden.

Next, we prospectively validated performance of Lung-CLiP in a cohort of 46 early-stage NSCLC patients and 48 risk-matched controls enrolled at an independent institution (Extended Data Fig. 9c). Stage-matched performance in the validation cohort was statistically similar to that observed in the training by AUC and sensitivity metrics (Fig. 4a–c, Extended Data Fig. 9d). Furthermore, specificity thresholds set in the training cohort performed similarly when applied to the controls in the validation cohort, indicating that Lung-CLiP scores are well calibrated (Extended Data Fig. 9e). For example, the 98% specificity threshold defined in the training cohort achieved a statistically similar specificity of 96% (95% CI: 89%−100%) in the validation cohort (55/56 training controls vs. 44/46 validation controls classified as negative, $P = 0.59$, Fisher's Exact Test). Finally, we explored the relationship between tumor volume and likelihood of detection by Lung-CLiP in the 103 NSCLC patients from the training and validation cohorts with MTV data available. We observed a strong correlation between MTV and sensitivity of Lung-CLiP (Fig. 4d, Extended Data Fig. 9f), with approximate sensitivities of 16% (95% CI: 4%−24%), 52% (95% CI: 32%−72%) and 80% (95% CI: 60%−96%) for 1 mL, 10 mL, and > 100 mL tumors, respectively.

Discussion

Here we describe a novel approach for noninvasive NSCLC screening that integrates improved molecular techniques with machine learning to predict the presence of NSCLC-derived cfDNA in a blood sample. Lung-CLiP achieves performance similar to tumor-informed ctDNA analysis without the need for tissue genotyping. Our approach differs from recent liquid biopsy studies that have attempted to develop pan-cancer screening assays[22–25]. Instead, we focused on NSCLC, allowing us to leverage lung cancer-specific features and to use control subjects who are at high-risk for developing the disease, a measure that reduces the likelihood that unrecognized confounders bias classification results. Additionally, unlike prior studies that did not perform validation or used cross-validation within the same case:control cohort[23–25], we employed an independent validation cohort that was prospectively enrolled at a different institution. This decreases the risk of model over-fitting leading to overly optimistic results[26]. Finally, although prior studies have also examined cfDNA fragmentation patterns[9,10,25], our study is unique in that we use this feature to aid in distinguishing tumor-derived mutations from their CH-derived counterparts.

We envision that one potential application of Lung-CLiP could be to serve as an initial screen in some of the ~95% of high-risk patients who are candidates for LDCT in the US, but who are not being screened due to a variety of issues including limited access and concerns with false positives[4,5]. Patients with positive Lung-CLiP tests would then be referred for LDCT. Although Lung-CLiP is less sensitive than LDCT, this hybrid approach could potentially increase the total number of patients screened and therefore the number of lives saved annually in the US from the current ~600 to closer to the projected maximum of ~12,000[27].

A striking observation in our study was the strong association of pre-treatment ctDNA levels with clinical outcomes in early-stage NSCLC, including within stage I patients. While these findings need to be validated, our results suggest that high pre-treatment ctDNA levels may reflect the presence of micro-metastatic disease and thus may allow identification of patients who would most benefit from neoadjuvant systemic therapy prior to surgery. Furthermore, it is possible that pre-treatment ctDNA measurements could be incorporated into NSCLC staging as well as enable real-time risk models that combine pre- and post-treatment variables to predict individualized patient outcomes[28].

Strengths of our study include the identification of molecular features differentiating tumor-derived and CH-derived cfDNA mutations, the use of risk-matched controls, and prospective validation. Our study also has a number of limitations. First, more patients need to be analyzed to fully establish performance characteristics of Lung-CLiP. Second, the majority of cases in our study were incidentally diagnosed lung cancers, and not identified by LDCT screening. Therefore, clinical screening sensitivity could be lower in a population setting and this should be prospectively evaluated. Third, we developed Lung-CLiP in a cohort mainly composed of smokers and therefore it is possible that performance could be worse in non-smokers.

In summary, we have developed an integrated genomic strategy that can detect a significant fraction of early-stage lung cancers using blood plasma. We envision that integration of Lung-CLiP with LDCT or other circulating biomarkers could further improve performance. Additionally, by modifying the at-risk populations considered and incorporating molecular features appropriate for other cancer types, we expect that it will be feasible to develop CLiP methods for a diverse range of malignancies.

Materials and Methods

Study design and patients

All samples analyzed in this study were collected with informed consent from subjects enrolled on Institutional Review Board-approved protocols that complied with all relevant ethical regulations at their respective centers, including Stanford University, MD Anderson Cancer Center, Mayo Clinic, Vanderbilt University Medical Center, and Massachusetts General Hospital.

Lung Cancer Patients: All patients had AJCC v7 stage I-III NSCLC and received curative-intent treatment with surgery or radiotherapy. This study consisted of two cohorts, a discovery cohort and a validation cohort. Clinical characteristics of patients in both cohorts are provided in Extended Data Fig. 4b. The discovery cohort consisted of two groups of patients: (1) tumor-informed NSCLC patients (Fig. 1 and Extended Data Fig. 5–7) and (2) Lung-CLiP training NSCLC cases (Fig. 2–3 and Extended Data Fig. 9a–b). These two groups consisted of lung cancer patients enrolled at Stanford University (n=80), Vanderbilt University (n=21), Mayo Clinic (n=14) and MD Anderson Cancer Center (n=7) between November of 2009 and July of 2018. The tumor-informed NSCLC cases consisted of 85 patients with matched tumor tissue available, the majority of which (67/85) were analyzed with all aspects of the improved CAPP-Seq workflow described in Fig. 1a. The Lung-CLiP training group was restricted only to patients analyzed with the improved workflow (n=104) and was studied for the tumor-naïve analyses in Fig. 2 and Extended Data Fig. 8, serving as the training group for the Lung-CLiP classifier (Fig. 3 and Extended Data Fig. 9a–b). Among the 104 Lung-CLiP training NSCLC cases, 67 overlap with the 85 patients in the tumor-informed group. After initial training of Lung-CLiP, NSCLC patients in the independent validation cohort (46 lung cancer cases; Fig. 4 and Extended Data Fig. 9c–d) were prospectively enrolled at Massachusetts General Hospital (MGH) between January and December of 2018.

Controls: The discovery cohort consisted of two separate control groups (Extended Data Fig. 4b). The first group consisted of 42 adult blood donors who were un-matched for risk (“low-risk controls”) and were only included in analyses presented in Fig. 2 and Extended Data Fig. 8. The second group consisted of 56 age-, sex- and smoking status-matched adults (“risk-matched controls”) who had negative low-dose computed tomography (LDCT) screening scans for lung cancer at Stanford University and served as the training group for the Lung-CLiP classifier (Fig. 2–3 and Extended Data Fig. 8, 9a–b). The validation cohort contained a third control group, comprised of 48 risk-matched adults undergoing LDCT screening at Massachusetts General Hospital, that were prospectively enrolled between January and December of 2018. This control group was only considered for the validation of the Lung-CLiP model (Fig. 4 and Extended Data Fig. 9c–d).

Blood collection and processing

Logistical considerations related to the prospective collection of the validation cohort required the use of STRECK blood collection tubes, while K_2EDTA collection tubes were used for the training cohort. Whole blood collected in K_2EDTA tubes was processed immediately or within 4 hours following storage at 4 °C. Whole blood collected in Cell-Free DNA BCT (STRECK) tubes was processed within 72 hours. K_2EDTA tubes were centrifuged once at 1,800 × g for 10 min and STRECK tubes were centrifuged twice at 1,600 × g for 10 min at room temperature. Following centrifugation, plasma was stored at −80°C in 1.8 ml aliquots until cfDNA isolation. Plasma-depleted whole blood was stored at −80°C for DNA isolation from leukocytes.

Our study design guards against pre-analytical variables such as blood collection tubes driving classification of cases versus controls because all samples within our training cohort (i.e. both cases and controls) were collected in K_2EDTA tubes while all samples within the validation cohort were collected in STRECK tubes. Nevertheless, to confirm that the type of collection tube does not confound the Lung-CLiP model we collected blood from three healthy donors in K_2EDTA and STRECK tubes and compared key metrics including Lung-CLiP classification, cfDNA mutation concordance, fragment size, cfDNA concentration, molecular recovery and error profiles and found that none of these were significantly affected by the type of collection tube used (Extended Data Fig. 10a–j).

Cell-free DNA was extracted from 2 to 16 mL of plasma (median of 3.25 ml for NSCLC patients and 3.91 ml for controls) using the QIAamp Circulating Nucleic Acid Kit (Qiagen) according to the manufacturer’s instructions. After isolation, cfDNA was quantified using the Qubit dsDNA High Sensitivity Kit (Thermo Fisher Scientific) and High Sensitivity NGS Fragment Analyzer (Agilent). Genomic DNA (gDNA) from matched plasma-depleted whole blood (i.e. “WBCs” or “leukocytes”) was extracted using the Qiagen DNeasy Blood and Tissue kit, quantified using Qubit dsDNA High Sensitivity Kit, and fragmented to a target size of 170 bp using Covaris S2 sonicator. Post-sonication, fragmented gDNA was purified using the QIAquick PCR Purification Kit (Qiagen). For cfDNA, a median of 38 ng (range 8–

85 ng; median of 38 ng and 40 ng in cases and controls, respectively) was input into library preparation. DNA input was scaled to control for high molecular weight DNA contamination. For gDNA from leukocytes, ≤ 100 ng of fragmented gDNA was input into library preparation.

Tumor tissue collection and processing

Tumor DNA was extracted from frozen biopsy samples using the Qiagen DNeasy Blood and Tissue kit or from FFPE biopsy samples using the Qiagen AllPrep DNA/RNA FFPE kit according to the manufacturer's instructions. Following extraction, DNA was quantified and fragmented in the same manner as gDNA from plasma depleted whole blood and ≤ 100 ng of sheared DNA was input into library preparation.

Library preparation and sequencing

We developed a new adapter schema, FLexible Error-correcting dupleX adapters ("FLEX adapters"), that de-couples the portion of the adapter containing the duplex molecular barcode (i.e. unique identifier or "UID") from the portion containing the sample barcode (Extended Data Fig. 3b). FLEX adapters utilize dual-index 8 bp sample barcodes (pairwise edit distances ≥ 5) and 6 bp error correcting UIDs (pairwise edit distances ≥ 3) with optimized GC content and sequence diversity. End repair, A-tailing, and adapter ligation are performed following the KAPA Hyper Prep Kit manufacturer's instructions with ligation performed overnight at 4°C. Adapter ligation is performed using a partial Y adapter containing a 6 bp UID and the T overhang required for ligation (Extended Data Fig. 3c). Following ligation, a bead cleanup is performed using SPRIselect magnetic beads (Beckman Coulter). Next, "grafting PCR" is performed to add dual-index 8 bp sample barcodes and the remaining adapter sequence necessary to make a functional Illumina sequencing library. Following another SPRI bead cleanup, universal PCR is performed. Additional details regarding the FLEX adapter design can be found in the Supplementary Methods.

Following library preparation, hybrid capture (SeqCap EZ Choice, NimbleGen) is performed. In this study we utilized a custom 355 kb NSCLC-focused panel targeting 255 genes recurrently mutated in lung cancer and 11 genes canonically associated with clonal hematopoiesis (Supplementary Table 1). Hybrid capture was performed according to the manufacturer's protocol, with the exception that hypotaurine (Sigma-Aldrich cat # H1384) was added to the hybrid capture reaction at a final working concentration of 5mM. All capture steps were conducted on a thermal cycler at 47 °C. Following enrichment, libraries were sequenced on an Illumina HiSeq4000 with 2×150 bp paired-end reads. Sequencing lane share was determined based on cfDNA input and the desired barcode family size. Median sequencing depths were 23,570x/5,012x (nominal/unique) for cases and 19,534x/4,075x for controls.

Sequencing data analysis and variant calling

Preprocessing and alignment: *FASTQ* files were demultiplexed using a custom pipeline in which read pairs were only considered if both 8 bp sample barcodes and 6 bp UIDs matched expected sequences following error-correction. Following demultiplexing, UIDs were removed and adapter read-through was trimmed from the 3' end of the reads using AfterQC to preserve short fragments[29]. Reads were aligned to the human reference genome (hg19) using BWA ALN[30].

Error suppression and variant calling: Molecular barcode-mediated error suppression and background polishing were performed as previously described[6]. To leverage the improved error profile afforded by capturing samples with the ROS scavenger hypotaurine, a background database built from 12 withheld healthy control plasma samples captured with hypotaurine was used for background polishing. Following error suppression, selector-wide single nucleotide variant (SNV) calling was performed as previously described using a custom variant calling algorithm optimized for the detection of low allele frequency variants from deep sequencing data[6]. This approach, termed “adaptive variant calling,” considers local and global variation in background error rates in order to determine position-specific variant calling thresholds within each sample. Variant calls were then further filtered as follows: (I) germline variants identified in WBC gDNA from any individual in the study at > 25% VAF were removed, (II) variants at low depth positions (< 50% of the median depth), and those in repeat, intronic, intergenic, or pseudogene regions were removed, (III) variants falling in regions with poor uniqueness or mappability were removed, (IV) variants with a population allele frequency > 0.1% in the gnomAD database[31] were removed, (V) recurrent background artifacts were removed using a blacklist specific to our targeted sequencing space derived from a database of 430 WBC gDNA samples. Following variant calling and filtering, additional filters were applied depending on the tissue compartment and analysis being performed (described below).

Tumor genotyping

Somatic variant calling in tumor tissue was performed as described in the prior section except that we required: (1) a minimum allele frequency threshold of 5%, (2) variants could not be present in the matched WBCs, (3) variants in intronic or intergenic were retained, and (4) variants in canonical clonal hematopoiesis genes other than *TP53* were removed.

Tumor-informed ctDNA detection

To query plasma for the presence of ctDNA using mutations identified in matched tumor tissue, we used our previously described Monte Carlo-based ctDNA detection index[6]. The ctDNA detection index threshold was set to achieve ≥ 98% specificity in 56 held-out control cfDNA samples from patients with negative LDCT scans analyzed using the same sequencing panel. In samples with detectable ctDNA the plasma VAF of each mutation tracked was adjusted based on the clonality and the copy number state of the mutations in the tumor (Extended Data Fig. 10l–n) as described in the Supplementary Methods. The ctDNA VAF for each sample was then calculated by averaging the VAFs of all tumor variants used for monitoring (including variants with 0 mutant reads in the sample).

In the tumor-informed CAPP-Seq analyses, the patient-specific analytical limit of detection (LOD) was determined as previously described[6]. Briefly, the LOD was estimated based on the binomial distribution, number of mutations tracked, and the number of cfDNA molecules sequenced (e.g. unique depth). In the present study, the LOD was defined as the lowest tumor fraction expected to yield 3 or more mutation-containing cfDNA molecules with 95%

confidence based on the binomial distribution, the number of mutations tracked, and the unique molecular depth. These patient-specific LODs were utilized in Fig. 1b–c and Extended Data Fig. 5c–d. Only patients with detectable ctDNA or an LOD < 0.01% were considered in Fig. 1c, with cases without detectable ctDNA classified as having ctDNA < 0.01%.

Capture panel design for customized CAPP-Seq

Whole-exome sequencing of tumor DNA and matched leukocyte DNA was performed for 17 patients using the SeqCap EZ Exome version 3.0 capture reagent (NimbleGen) according to the manufacturer's protocol. Sequencing data were demultiplexed and mapped as described above and duplicate reads were removed using 'samtools rmdup'. Single-nucleotide variants were called using VarScan2[32], Mutect[33] and Strelka[34]. Variants called by ≥ 2 callers were then further filtered requiring: (i) VAF ≥ 5%, (ii) ≥ 30X positional depth in both tumor and germline, (iii) 0 germline reads, (iv) a population allele frequency ≤ 0.1% in the gnomAD database[31], and removing variants lying in repeat, intronic, intergenic, or pseudogene regions. Custom capture panels (SeqCap EZ Choice, NimbleGen) were then designed, each targeting the union of mutations from 5–7 patients and ranging in size from 212–487 kb. Tumor and matched leukocyte sequencing libraries from each patient were re-captured using these custom panels and tumor variants were re-called from the targeted sequencing data using the standard CAPP-Seq pipeline. These final variant lists, targeting a median of 68 mutations per patient (range 7–543), were then used for ctDNA detection.

ctDNA detection for customized CAPP-Seq

To query for the presence of ctDNA using custom CAPP-Seq panels, we applied the same Monte Carlo-based sampling approach[6] used for standard CAPP-Seq tumor-informed detection to two different subsets of molecules: (i) cfDNA molecules for which both strands of the original cfDNA duplex were observed and (ii) cfDNA molecules in the ctDNA enriched regions (Fig. 1e). We then combined these two *P*-values using Fisher's method. The ctDNA detection index threshold then was set to achieve ≥ 98% specificity in 24 healthy control cfDNA samples analyzed using the same sequencing panel.

Cancer cell fraction analysis

To determine the clonality of mutations identified in tumor samples, ABSOLUTE was used as previously described[35] to estimate the fraction of tumor cells harboring each somatic mutation (i.e. cancer cell fraction, CCF). Genome-wide segmented copy number calls (see "Detection of genome-wide copy number variation from targeted sequencing" section of Supplementary Methods) and the positions and VAFs of point mutations were used as input. Clonal mutations were defined as those for which the upper bound of the CCF confidence interval was > 0.95, while mutations with CCF estimates below this threshold were defined as subclonal. If only 1 mutation was identified in a tumor sample this mutation was considered to be clonal as it was not possible to obtain a CCF estimate.

ctDNA fragment size analysis

To compare the size distribution of tumor-derived and non-mutant cfDNA molecules we queried the plasma for cfDNA molecules overlapping the genomic positions of mutations identified in matched tumor samples. We then extracted the cfDNA fragment size (TLEN field in SAM Spec v1.6) of each molecule containing a tumor-derived mutation (i.e. "mutant molecules" or "ctDNA") and every non-mutant molecule spanning the same genomic position in the same individual. We then pooled mutant and non-mutant fragment lengths across all positions to generate the fragment size distributions depicted in Fig. 1d. We applied the same methodology to cfDNA mutations identified following tumor-naïve variant calling to generate the "CH" and "Tumor-adjudicated" mutation fragment size distributions depicted in Fig. 2i.

To determine what fragment size windows were enriched for ctDNA, we calculated the fraction of all mutant and non-mutant molecules falling in a 5 bp sliding window using the rollapply function in R (zoo package). We then calculated the relative enrichment of mutant vs. non-mutant molecules (i.e. "ctDNA enrichment") for every cfDNA fragment size between 50–500 bp (Fig. 1e).

Clinical correlates of ctDNA detection

Metabolic tumor volume was determined using whole body [18F] FDG positron emission tomography (PET)-CT scans. Percent ground glass opacity (GGO) and the presence of necrosis were determined using pretreatment imaging with chest computed tomography (CT) by a thoracic radiologist. GGO was defined by the presence of hazy, increased opacity of the lung with preservation of the bronchial and vascular margins[11]. Percent GGO was determined by examining the entire volume of the lesion on axial, sagittal, and coronal reconstructions with percent GOO in the entire tumor quantified and rounded to the nearest quartile. Multivariable linear regression was performed to associate the predictor variables (with stage and histology as categorical variables and MTV as a continuous variable) with mean ctDNA VAF (as the continuous dependent variable; Extended Data Fig. 6c). For patients without detectable ctDNA, a VAF of 0.001% was used. MTV and mean ctDNA VAF were log transformed to produce normally distributed data. The linear regression model was statistically significant ($P < 2.2 \times 10^{-16}$) and the residuals of the model were normally distributed as determined by the Shapiro-Wilk normality test.

We considered the following survival endpoints: (1) freedom from recurrence (radiographic or biopsy proven recurrence with censoring of non-cancer deaths), (2) freedom from metastasis (radiographic or biopsy proven metastasis to a distant organ or the contralateral lung with censoring of non-cancer deaths), (3) recurrence-free survival (radiographic or biopsy proven recurrence or death from any cause), (4) metastasis -free survival (radiographic or biopsy proven metastasis to a distant organ or the contralateral lung or death from any cause), (5) overall survival (death from any cause). Median follow up for the cohort was 30.1 months and 78/85 (92%) of patients were treated surgically (Supplementary Table 3). Patients without events were censored at last radiographic follow-up. Survival probabilities were estimated using the Kaplan-Meier method and survival of groups was compared using a two-sided log-rank test. Regression analysis was performed by Cox proportional hazards modeling, *P*-values were assessed using the log-likelihood test, and all *P*-values were two-sided. For regression analyses, log-transformed mean VAF and tumor volume measurements were used; log transformation was performed to produce normally distributed data. For patients without detectable ctDNA, a VAF of 0.001% was used. Continuous variables were standardized to enable comparison of hazard ratios and 95% confidence intervals using Cox models. The proportional hazard assumption of the Cox proportional hazard models (Fig. 1k and Extended Data Fig. 7e) was satisfied. We tested the proportional hazard assumption by Schoenfeld residuals and the *P*-values for the test of proportional hazards for each individual covariate are provided in Supplementary Table 11.

Characterization of clonal hematopoiesis in cfDNA and WBCs

To characterize clonal hematopoiesis (CH) in the cfDNA and WBC compartments (Fig. 2, Extended Data Fig. 8) we began with variants called as described in the "Error suppression and variant calling" section of the Methods with the following additional filters: (1) only nonsynonymous mutations were considered except for the positive selection analysis (Extended Data Fig. 8f) and the mutational signature analysis (Fig. 2g) for which synonymous mutations were also considered, (2) mutations were rescued from blacklisting if they were in the following 12 genes canonically associated with CH: *ASXL1, PPM1D, DNMT3A, TET2, GNB1, CBL, JAK2, STAT3, GNAS, MYD88, SF3B1, TP53*, and (3) mutations in canonical lung cancer driver genes[36] were rescued from blacklisting if they had been observed in ≥ 10 COSMIC lung cancer cases (CosmicGenomeScreens v85).

Using matched white blood cell (WBC) sequencing, mutations identified in the cfDNA were labeled as WBC-, WBC+, or WBC-undetermined as follows:

i. A mutation was considered WBC+ if it was above background in matched WBCs as assessed using the same Monte Carlo approach used for tumor informed ctDNA detection and requiring a detection index *P*-value < 0.05.

ii. A mutation was considered WBC- if there were 0 supporting reads in the matched WBC DNA **and** there was sufficient depth in the matched WBC DNA to identify the mutation given the VAF observed in plasma. Specifically, a mutation was only labeled WBC- if the probability of observing ε 1 supporting read in the WBCs was > 95% given the VAF of the variant in the cfDNA and the positional depth in the WBCs.

iii. A mutation was considered WBC-undetermined if there were > 0 supporting reads in the WBCs but the detection index *P*-value was ≥ 0.05 (i.e. mutation was not significantly above background in WBCs) **or** if there were 0 supporting reads but the probability of observing the mutation in the matched WBCs was ≤ 95% given the VAF of the variant in the cfDNA and positional depth in the WBCs.

Only mutations identified *de novo* in the cfDNA for which presence in the matched WBCs could be confidently assessed (labeled as WBC- or WBC+) were considered for all the analyses in Fig. 2 and Extended Data Fig. 8 with the following exceptions:

i. For Extended Data Fig. 8b, mutations identified *de novo* from WBCs were also considered.

ii. For the analysis comparing VAFs of mutations found in cfDNA and WBCs (Fig. 2e), mutations called *de novo* in either compartment (cfDNA or WBCs) were considered as long as the presence or absence of the alteration could be confidently assessed in both tissue compartments, as detailed above. Therefore, mutations identified de novo in WBCs were labeled as cfDNA-, cfDNA+, or cfDNA-undetermined in the same manner that WBC support was determined for cfDNA mutations (see above). All mutations identified in either compartment are provided in Supplementary Tables 7–8.

Positive selection analysis of CH-derived cfDNA mutations

Positive selection analysis was carried out on all synonymous and nonsynonymous WBC+ and WBC- cfDNA mutations using the dNdScv R package[37] with a modification to account for the fraction of a given gene covered by our sequencing panel (see Data and code availability section). Genes were considered under positive selection for nonsynonymous mutations if the dNdScv-reported Q-value for all substitution types was < 0.05. All genes meeting this threshold are displayed in Extended Data Fig. 8f.

Mutational signature analysis of WBC+ and WBC- cfDNA mutations

The contribution of known mutational processes to the mutations we observed in cfDNA was assessed with the deconstructSigs R package[38] using the COSMIC signature set (v2). Due to the limited number of mutations per individual, mutations were pooled across individuals to evaluate mutational signatures present in WBC+ and WBC- compartments for a given comparison (e.g. WBC+ mutations in patients vs. WBC+ mutations in controls). Signatures recurrently observed across groups are displayed in Fig. 2g. To assess the statistical significance of differences in the contribution of Signature 4 (smoking) to different sets of mutations, we performed 1,000 permutations per comparison of interest in which mutation labels were scrambled and mutational signature contributions were recalculated with deconstructSigs. For each permutation, the difference in Signature 4 contributions between the two mutation groups was computed to generate a null distribution, and an empirical *P*-value was determined by comparing the observed difference in Signature 4 between true mutation groups to the null distribution. To correct for mutation sets that had imbalanced label counts due to differences in cohort size (i.e. different numbers of mutations in the groups being compared), we randomly down-sampled the number of mutations to the less-represented label's total in each iteration before recalculating the mutational signature contributions.

Data and code availability

Anonymized clinical and demographic data on the lung cancer cases and non-cancer controls considered in this study, as well as cfDNA metrics, cfDNA and WBC somatic mutation data, Lung-CLiP scores, and other relevant data are provided in the Supplementary Tables. The detailed patient-level genomic features used as input for the Lung-CLiP model (including genome wide somatic copy number alteration data and somatic mutation genotyping data with all the associated features considered in the Lung-CLiP model), along with code for the Lung-CLiP classification model, the in silico simulation of the CAPP-Seq molecular biology workflow, and the modified dNdScv R functions[37] (accounting for the fraction of a given gene covered by our sequencing panel) can be found at http://clip.stanford.edu. This website provides users with the code and data used for the training and validation of the Lung-CLiP model and the *in silico* simulation of the CAPP-Seq molecular biology workflow, allowing for reproduction of our results and figures. Due to restrictions related to dissemination of germline sequence information included in the informed consent forms used to enroll study subjects, we are unable to provide access to raw sequencing data. Reasonable requests for additional data will be reviewed by the senior authors to determine whether they can be fulfilled in accordance with these privacy restrictions.

To assign each mutation a score reflecting the likelihood it resulted from smoking-associated mutational processes (Fig. 2h), we considered the trinucleotide context and base substitution for the mutation and then extracted the weight for that context from the COSMIC Signature 4 vector as provided by deconstructSigs.

Droplet digital PCR

We performed an orthogonal validation of 15 WBC+ cfDNA mutations observed in a subset of patients and controls using droplet digital PCR (ddPCR). ddPCR was performed on a Bio-Rad QX200 instrument as previously described[39] using reagents, primers, and probes obtained from Bio-Rad. We validated four private mutations, as well as two recurrent hotspot mutations in *DNMT3A* and *JAK2* that were observed in 11 cfDNA samples. We found that 100% (15/15) of the mutations we tested validated by ddPCR in both the cfDNA and WBC gDNA compartments and that VAFs quantified by CAPP-Seq and ddPCR were significantly correlated (Extended Data Fig. 10k).

Detection of genome-wide copy number variation from targeted sequencing

To identify copy number variants (CNVs), we utilized both the on- and off-target reads from CAPP-Seq. Briefly, each library in the CAPP-Seq workflow typically receives ~30–60 million paired-end reads. These reads are mapped to the human genome (build GRCh37/hg19), with ~60–80% of reads falling in the targeted genomic coordinates ("on-target reads"). The remaining 20–40% of reads predominantly map to the remainder of the human genome ("off-target reads"). To combine the high-depth data in our targeted sequencing space with the low-pass data in the off-target space, we treat each of these sets of reads separately, followed by statistical integration (described in detail in the "Detection of genome-wide copy number variation from targeted sequencing" section of the Supplementary Methods).

Lung-CLiP model

The Lung-CLiP model is an ensemble classification framework integrating the outputs of two constituent SNV and CNV models using five different classification rules, *5-nearest neighbor (5NN)*, *3NN*, *naïve Bayes*, *logistic regression* and *decision tree*. Detailed descriptions of the SNV model, CNV model, and integrated Lung-CLiP ensemble classifier and the features used in each model are described in detail in the Supplementary Methods. Briefly, for the SNV model we developed a statistical model to distinguish cfDNA mutations observed in patients from those observed in controls. Within this model we leverage a *semi-supervised learning* framework in which an elastic net logistic regression model is trained to distinguish tumor-adjudicated variants from non-adjudicated variants ('tumor-adjudicated model') in the subset of patients with matched tumors. This tumor-adjudicated model is used to label variants from patients without matched tumor samples. The SNV model is then used to assign scores to all variants in patients and controls using the labels assigned by the semi-supervised tumor-adjudicated model. After variant scores have been assigned, we perform "*Patient SNV Featurization*" to summarize the variant scores in each sample. These summary scores are then used in a final elastic net logistic regression model trained to distinguish patients from controls. All these steps were performed in a nested patient-level leave-one-out framework in the training cohort.

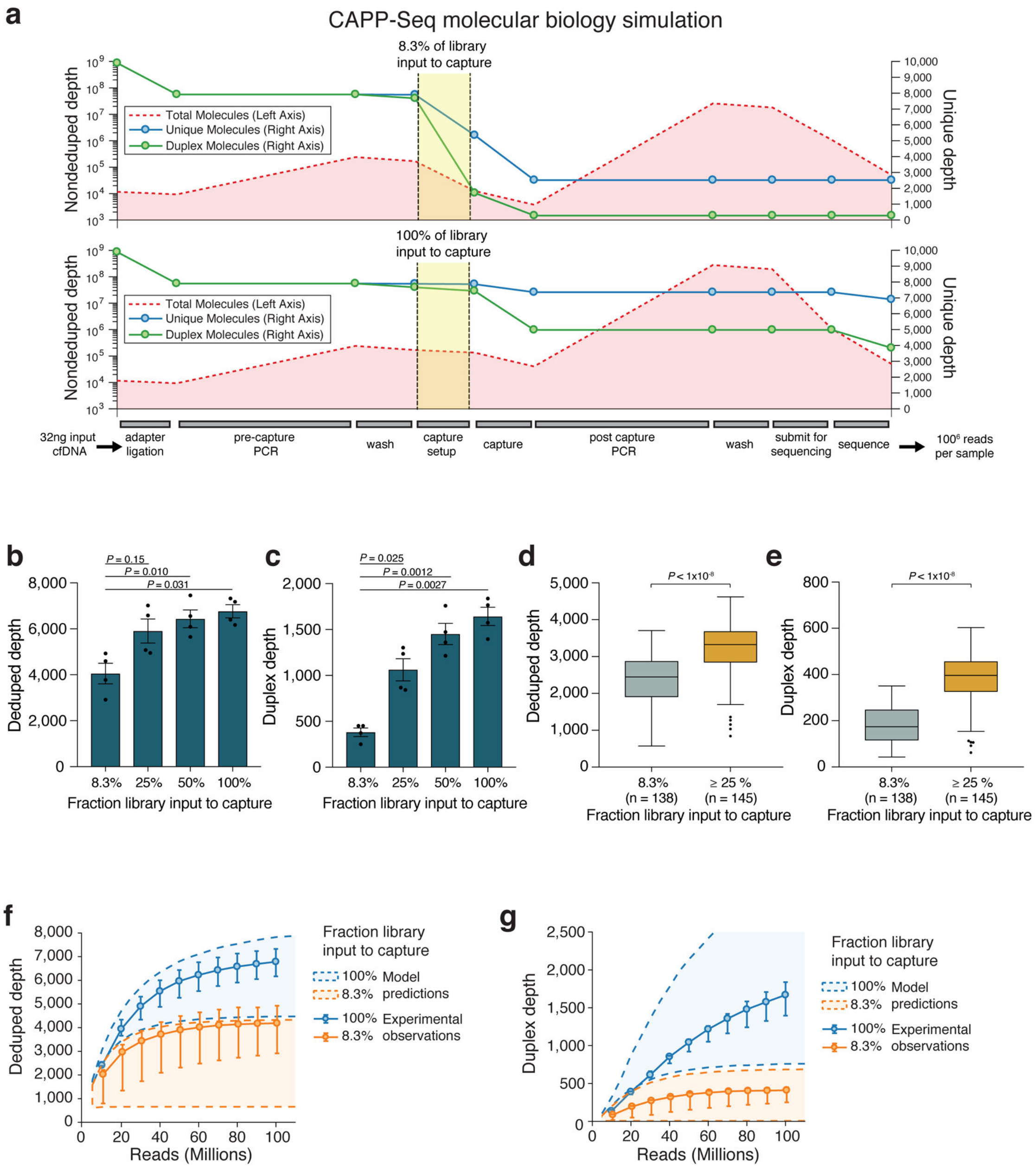
The CNV model enumerates altered genomic regions using two annotation lists: (1) a set of uniformly distributed 5 MB windows across the genome, and (2) recurrently altered regions identified by running GISTIC2.0[40] on 1,017 TCGA NSCLC cases (i.e. “hotspot regions”). Following filtering steps described in the Supplementary Methods, the number of 5 MB regions and GISTIC “hotspot” regions are used as features in the copy number model alongside a third feature which captures whether there is enrichment for regions known to be recurrently copy number altered in NSCLC (i.e. GISTIC) as opposed to uniform 5 MB bins.

In an exploratory analysis we combined the training and validation cohorts to examine whether sequencing depth or related metrics may influence detection by Lung-CLiP. We found that cfDNA input, plasma volume input, and unique sequencing depth were not significantly associated with the sensitivity of Lung-CLiP (Extended Data. Fig. 9g–i).

Statistical analysis

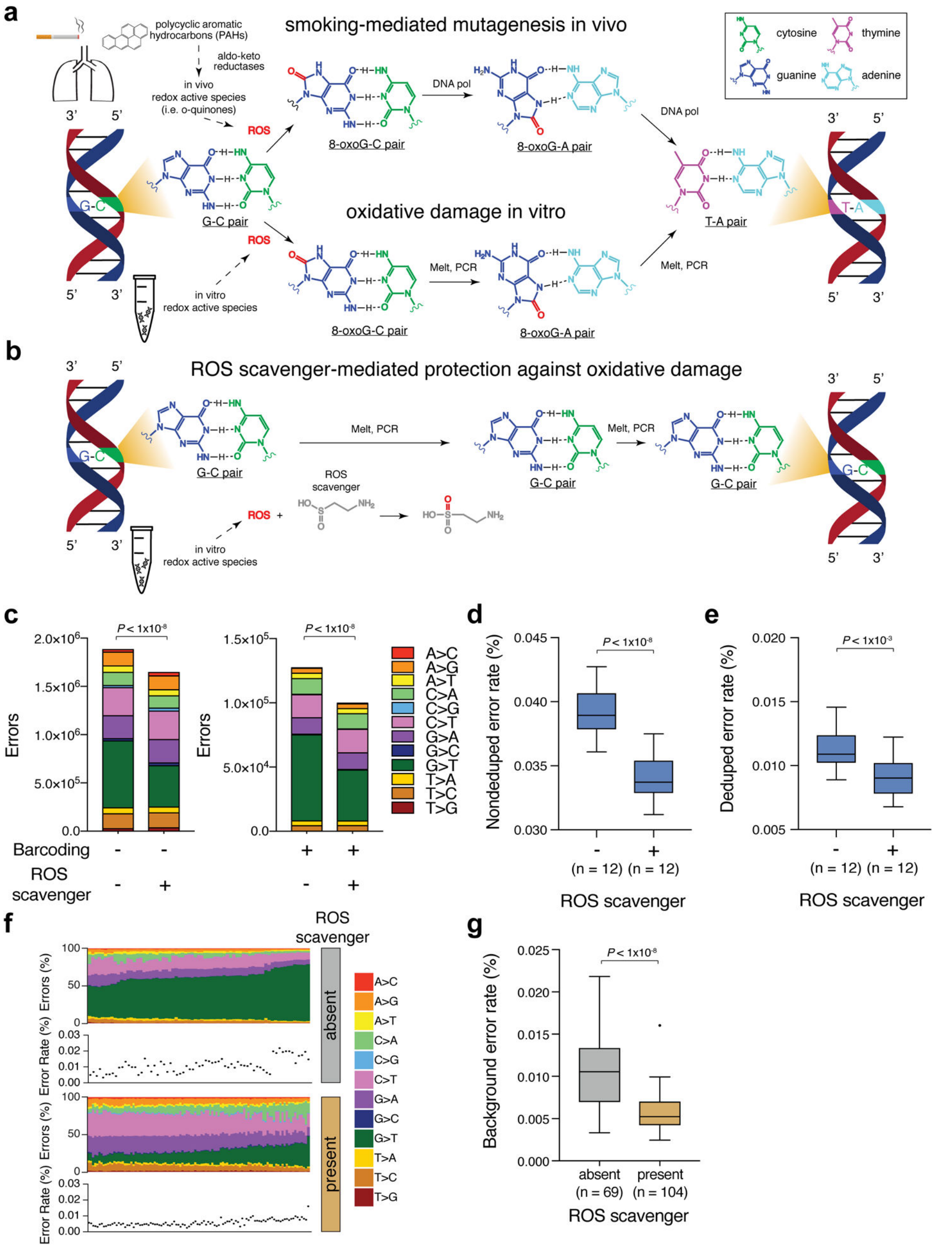
Statistical analyses were performed in R (version 3.4.0 and 3.5.2), MATLAB (R2018a) and GraphPadPrism7 (version 8.3.0). Statistical tests used throughout the manuscript include the Wilcoxon rank-sum test, paired t-test, Fisher’s Exact Test, Pearson correlation, Spearman correlation and Cox proportional hazards model. Unless otherwise specified, the sample size (n) denoted in the text, figure panels and figure legends refer to biologically independent individuals or mutations. Unless otherwise specified, all statistical tests were two-sided, no adjustments were made for multiple comparisons when performing grouped comparisons, and analyses for significant differences between two groups were conducted using the Wilcoxon rank-sum test. Unless otherwise specified, in violin plots the horizontal dashed lines denote the median and interquartile range, and in box plots the boxes capture the interquartile range, the center line denotes the median and the whiskers depict the extrema. When assessing correlation by Pearson or Spearman correlation, statistical significance was assessed using t-statistics. Survival probabilities were estimated using the Kaplan-Meier method and survival of groups of patients were compared using the log-rank test. Regression analysis was performed by Cox proportional hazards modeling, *P*-values were assessed using the log-likelihood test, and all *P*-values were two-sided. Multivariable analysis of clinical correlates of ctDNA burden was performed by linear regression. The Lung-CLiP classification framework employs the R packages glmnet, caret, ETC, pROC, survival, optparse and MASS. Confidence intervals for sensitivity, specificity, and AUC estimates of Lung-CLiP were generated by 1,000 bootstrap re-samplings of the Lung-CLiP classification scores in the training and validation cohorts. Details of the statistical models used in the Lung-CLiP classification framework and the *in silico* simulation of the CAPP-Seq molecular biology workflow can be found in the Supplementary Methods. A power analysis was performed to determine an appropriate size for the Lung-CLiP validation cohort. Assuming a specificity of 98% as determined in the training cohort, we calculated that 48 controls would have 80% power to detect that the true specificity is ≥ 90% (1 arm binomial test with one sided alpha = 0.05). Statistical significance for tumor-informed ctDNA detection was determined with our previously-described Monte Carlo-based ctDNA detection index[6] as described in the Methods. Statistical significance of the smoking mutational signature contribution to select mutation sets was performed by permuting SNV labels as described in the Methods. Positive selection analysis was carried out on WBC+ and WBC- cfDNA mutations using the dNdScv R package[37] as described in the Methods.

Extended Data



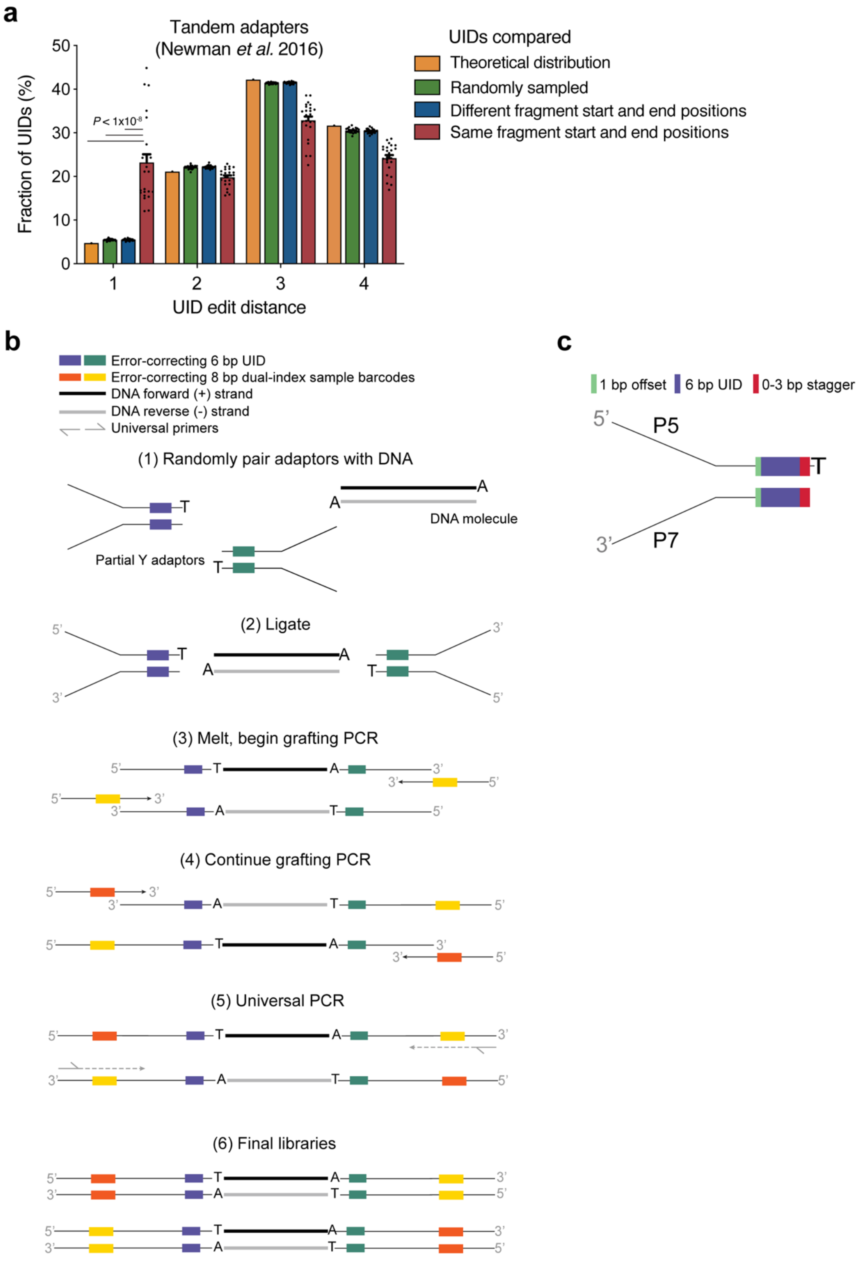
Extended Data Figure 1. Development and experimental validation of an *in silico* simulation of the CAPP-Seq molecular biology workflow.

(**a**) The fraction of original unique (blue line) and duplex (green line) cfDNA molecules (‘Unique depth’, right axis) and total molecules including PCR duplicates (‘Nondeduped depth’, left axis) at each step in the CAPP-Seq molecular biology workflow were tracked using an *in silico* model based on random binomial sampling. In this model, only on-target molecules are considered, with both individual DNA strands from original DNA duplexes tracked. Two simulations are shown, with 8.3% (top) and 100% (bottom) of amplified sequencing library input into the hybridization reaction for target enrichment. Additional details on the model are provided in the Supplementary Methods. (**b-c**) Empirical validation of simulation models. Comparison of median unique (**b**) de-duplicated (i.e. "deduped") and (**c**) duplex depths recovered by sequencing following input of different fractions of sequencing library into the hybrid capture reaction. A total of 32 ng of cfDNA from each of 4 healthy adults was used as input in each condition and each sample was downsampled to 100 million sequencing reads prior to barcode-deduplication to facilitate comparison. Comparisons were performed with a paired two-sided t-test. (**d-e**) Comparison of (**d**) deduped and (**e**) duplex sequencing depths achieved following input of 8.3% (n=138 cfDNA samples) compared to ≥ 25% (n=145 cfDNA samples) of each sequencing library into the hybrid capture reaction. All samples had 32 ng of cfDNA as input to library preparation and were downsampled to 25 million reads prior to barcode-deduplication to facilitate comparison. In box plots the center line denotes the median, the box contains the interquartile range, and the whiskers denote the extrema that are no more than 1.5 × IQR from the edge of the box (Tukey style). (**f-g**) Comparison of deduped (**f**) and duplex (**g**) sequencing depths predicted by the model to that observed experimentally when 8.3% vs. 100% of a sequencing library is input into the hybrid capture reaction. A range of capture efficiencies (7.5 – 75% hybrid capture efficiency) were considered in the simulation, where the confidence envelope denotes the resultant range of model predictions. The experimental data depicted in panels **b-c** (n=4 cfDNA samples per capture condition) was downsampled prior to barcode deduplication to enable comparisons across different sequencing read yields (x-axis). Dots denote the median and error bars denote the minimum and maximum.



Extended Data Figure 2. The ROS scavenger hypotaurine reduces oxidative damage arising *in vitro*.

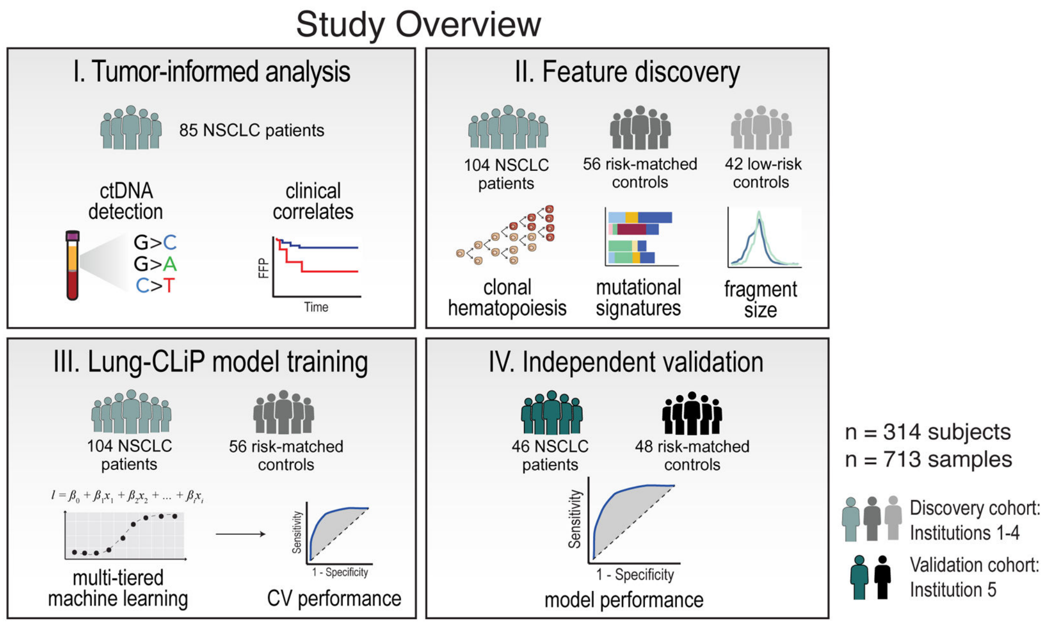
(**a**) Diagram illustrating the chemical mechanism by which carcinogens in cigarette smoke *in vivo* (top) or reactive oxygen species (ROS) *in vitro* (bottom) cause damage to DNA leading to the generation of 8-oxoguanine, which subsequently results in the generation of G>T transversions. (**b**) Diagram illustrating the proposed mechanism by which the addition of a ROS scavenger reduces oxidative damage-derived G>T artifacts *in vitro*. (**c**) Comparison of base substitution distributions in healthy control cfDNA samples (n=12 individuals) captured with and without the ROS scavenger hypotaurine present in the hybrid capture reaction. The number of errors that are G>T transversions was compared using a paired two-sided t-test ($P < 1\times10^{-8}$). (**d-e**) Aggregate selector-wide nondeduped (**d**) and deduped (**e**) background error rates summarizing results in panel **c**. Grouped comparisons were performed with a paired two-sided t-test. (**f**) Comparison of selector-wide error rates and base substitution distributions across two cohorts of healthy controls, where cfDNA samples were profiled with ("present," bottom, n=104) or without ("absent," top, n=69) the ROS scavenger hypotaurine present in the hybrid capture reaction. (**g**) Aggregate selector-wide error rates summarizing results from panel **f**. In box plots the center line denotes the median, the box contains the interquartile range, and the whiskers denote the extrema that are no more than 1.5 × IQR from the edge of the box (Tukey style).



Extended Data Figure 3. Rationale for and overview of dual-index duplex adapters with error-correcting barcodes (i.e. FLEX adapters).

(**a**) An excess of molecular barcodes (i.e. unique identifier or “UIDs”) differing by 1 bp in cfDNA molecules with the same the start and end positions indicates that sequencing errors in UIDs can create erroneous UID families. Depicted is the expected and observed distribution of barcode Hamming edit distances (“UID edit distance”) when comparing UIDs from different groups of barcode-deduped (i.e. unique) cfDNA molecules sequenced using our previously described tandem adapters[6]. Tandem adapters utilize random 4-mer UIDs, resulting in 256 distinct UIDs that cannot be error corrected. The theoretical distribution of UID edit distances across all 256 UIDs is shown in orange (i.e. the fraction of UIDs that differ from one another by 1, 2, 3, and 4 bp). The green, red and blue bars represent the distribution of UID edit distances observed in healthy control cfDNA samples sequenced with tandem adapters (n=24 individuals). Green indicates randomly sampled UIDs, blue indicates UIDs from cfDNA molecules with different genomic start/end positions, and red indicates cfDNA molecules sharing the same start/end positions. UIDs differing by only one base are significantly overrepresented when comparing cfDNA molecules with the same start/end position (red bars) to each of the other UID distributions, suggesting that 1 bp errors are erroneously creating new UID families. Group comparisons were performed with a paired two-sided t-test, except when comparing to the theoretical distribution, for which an un-paired two-sided t-test was used ($P < 1\times10^{-8}$). Bars denote the mean and error bars denote the standard error. (**b**) Schematic overview of custom FLexible Error-correcting dupleX (‘FLEX’) sequencing adapters, enabling independent tailoring of UID diversity and multiplexing capacity. Shown is an initial DNA molecule to which ‘partial Y adapters’ containing duplex UIDs are ligated (1–2). Next, the two molecules derived after one round of ‘grafting PCR’ (which adds the first of two sample barcodes) are shown (3). This is followed by additional rounds of grafting PCR which add the second sample barcode and continues to amplify the library (4). Following grafting PCR, a magnetic bead cleanup is performed (not shown) which is followed by universal PCR (5), after which final sequencing libraries compatible with Illumina sequencers are shown (6). Dual index sample barcodes types are indicated in yellow (‘index 1’ or ‘i7’) and orange (‘index 2’ or ‘i5’) and UIDs are indicated by purple/green blocks. (**c**) Diagram depicting a detailed view of the ‘partial Y adapters’ used for initial ligation to cfDNA. The adapters contain a ‘1 bp offset’ indicated in green, followed by a 6 bp error correcting UID indicated in purple (Hamming edit distances ≥ 3), followed by 0–3 ‘stagger’ bases indicated in red, followed by a 3’ ‘T-overhang’ for ligation. The 0–3 bp stagger bases increase sequence complexity early in the sequencing reads to obviate the need for PhiX (used for spectral diversity). Additional details on the FLEX adapters are provided in the Supplementary Methods.

a



b

Lung-CLiP discovery cohort

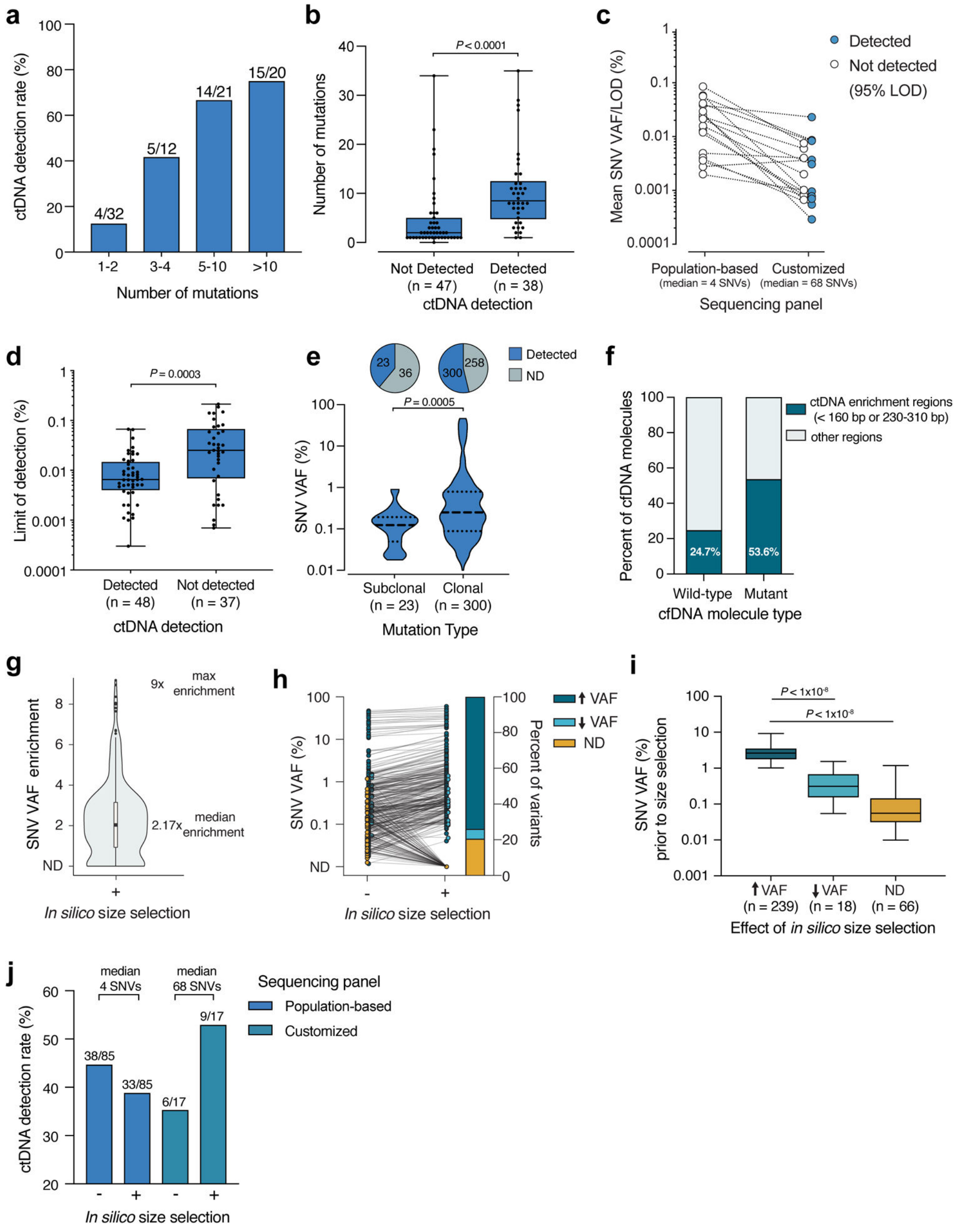
Parameter	Tumor-informed NSCLC patients n = 85	CLiP training NSCLC patients n = 104	Risk-matched controls n = 56	Low-risk controls[b] n = 42	*P*-value[cd]
Gender					0.87
Male	50 (59%)	63 (61%)	35 (62%)	22 (52%)	
Female	35 (41%)	41 (39%)	21 (38%)	20 (48%)	
Age (years)	70 (42-87)	70 (42-87)	69 (54-83)	45 (22-70)	0.76
Smoking					
Yes	64 (75%)	83 (80%)	56 (100%)	1 (2%)	
No	21 (25%)	21 (20%)		41 (98%)	
Pack-years	28 (0-135)	30 (0-135)	40 (20-136)	-	0.09
Stage[a]					
IA	15 (18%)	21 (20%)	-	-	
IB	33 (39%)	28 (27%)	-	-	
IIA	9 (11%)	12 (12%)	-	-	
IIB	12 (14%)	16 (15%)	-	-	
IIIA	12 (14%)	17 (16%)	-	-	
IIIB	4 (5%)	10 (10%)	-	-	
Histology					
Adenocarcinoma	63 (74%)	71 (68%)	-	-	
Squamous	18 (21%)	23 (22%)	-	-	
NOS	2 (2%)	7 (7%)	-	-	
Large Cell	2 (2%)	3 (3%)	-	-	
Institution					
Stanford	53 (62%)	76 (73%)	56 (100%)	42 (100%)	
Vanderbilt	18 (21%)	21 (20%)			
Mayo Clinic	14 (16%)	-			
MD Anderson	-	7 (7%)	-		

Lung-CLiP validation cohort

Parameter	NSCLC patients n = 46	Risk-matched controls n = 48	*P*-value[c]
Gender			0.31
Male	21 (46%)	28 (56%)	
Female	25 (54%)	20 (44%)	
Age (years)	69 (52-83)	66 (55-78)	0.30
Smoking			
Yes	46 (100%)	48 (100%)	
No	-	-	
Pack-years	40 (20-165)	39 (23-132)	0.81
Stage[a]			
IA	22 (48%)	-	
IB	10 (22%)	-	
IIA	-	-	
IIB	9 (20%)	-	
IIIA	2 (4%)	-	
IIIB	3 (7%)	-	
Histology			
Adenocarcinoma	36 (78%)	-	
Squamous	10 (22%)	-	
Institution			
Harvard	46 (100%)	48 (100%)	

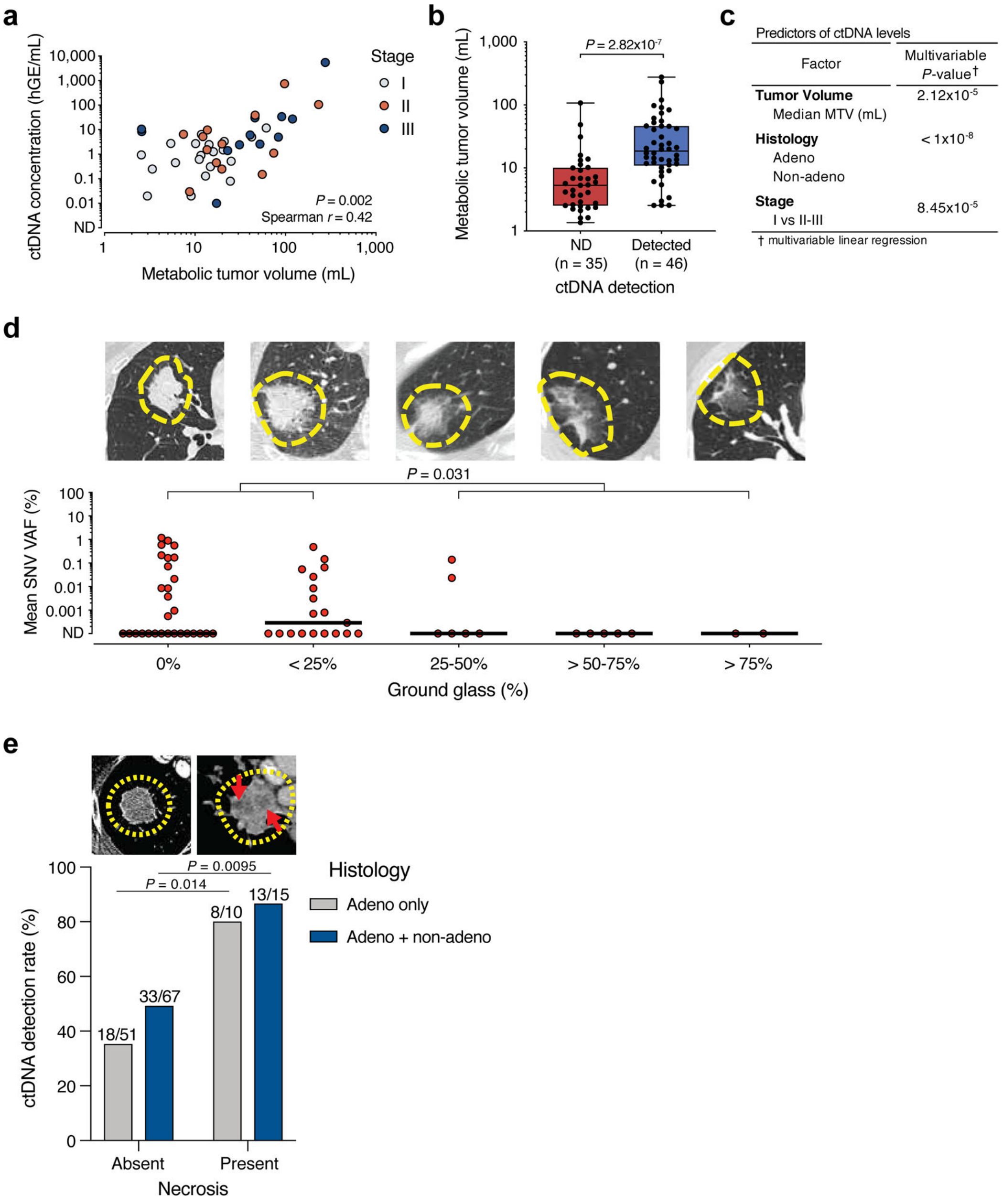
Extended Data Figure 4. Study and cohort overview.

(**a**) Study Overview. (**b**) Clinical and demographic information pertaining to the NSCLC patient and non-cancer control cohorts considered in this study. For categorical variable, the count is provided with the percent of the cohort in parentheses. For continuous variables, the median value is provided with the range of values in parentheses. NOS = not otherwise specified, a = AJCC v7 staging, b = Low-risk controls were considered for feature discovery and CH analysis only and were not used for Lung-CLiP model training, c = Sex was compared with a two-sided Fisher's Exact Test and continuous variables (age and pack-years) were compared with an un-paired two-sided t-test, d = Lung CLiP NSCLC patients and risk-matched controls were compared.



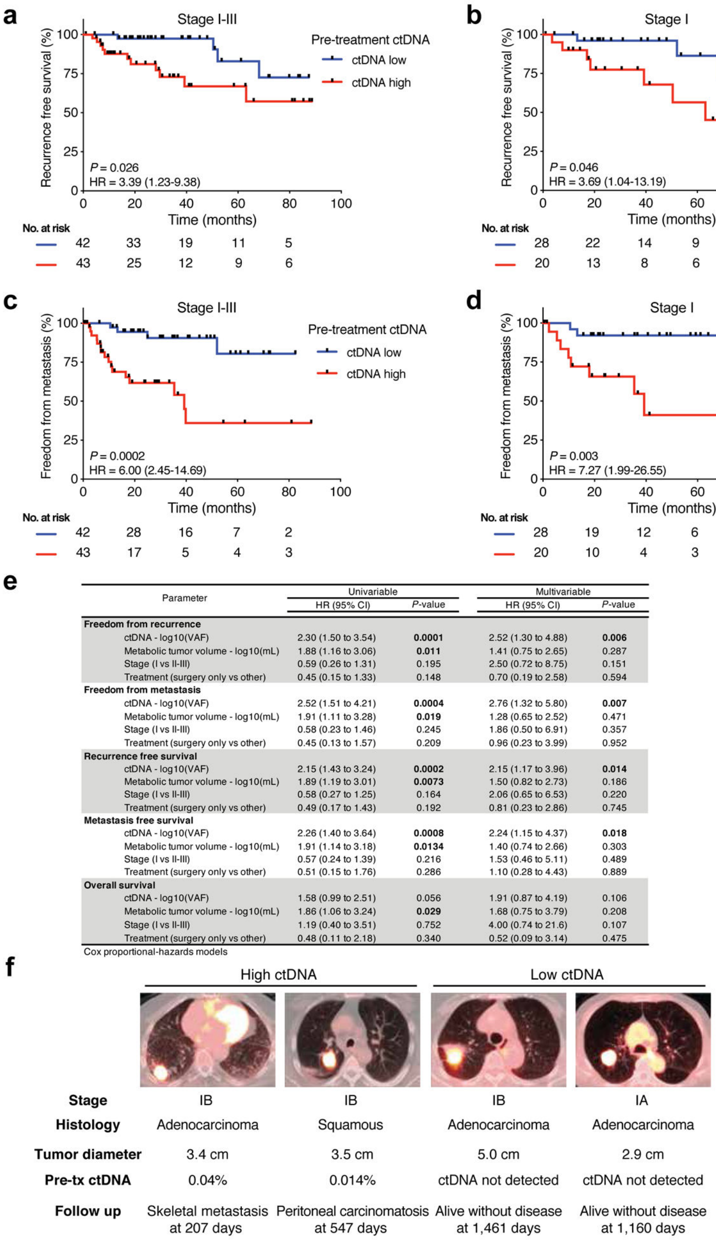
Extended Data Figure 5. Biological determinants of tumor-informed ctDNA detection.

(**a**) Association between tumor-informed ctDNA detection and the number of mutations tracked using the population-based lung cancer-focused CAPP-Seq panel. All patients were considered and binned by the number of mutations identified in matched tumor biopsy samples. (**b**) Association between the number of mutations identified in matched tumor samples and tumor-informed ctDNA detection using the population-based lung cancer-focused CAPP-Seq panel. (**c**) ctDNA detection statistics in 17 early-stage NSCLC patients profiled both with the population-based lung cancer-focused CAPP-Seq panel (left), and customized capture panels designed using tumor exome sequencing data (right). While all 17 patients were undetectable using the population-based method, 10 (59%) were detected using customized panels. For samples without detectable ctDNA (open circles), the corresponding patient-specific analytical limit of detection (LOD) is shown. For patients with detectable ctDNA, the mean variant allele frequency (VAF) observed across all tracked mutations is depicted (blue circles). (**d**) Comparison of the patient-specific analytical limit of detection (LOD) in patients with and without detectable ctDNA using tumor-informed CAPP-Seq. LOD was determined based on the binomial distribution, number of mutations tracked, and the number of cfDNA molecules sequenced (e.g. unique depth). The LOD from patients sequenced with the population-based lung cancer-focused CAPP-Seq panel only (n=68) and patients sequenced with customized capture panels designed using tumor exome sequencing data (n=17 patients) are displayed. (**e**) Detection of clonal and subclonal SNVs in cfDNA. The fraction of all clonal and subclonal SNVs detected in plasma are depicted in pie charts (two-sided Fisher's Exact Test, $P = 0.039$) and the VAFs of clonal and subclonal SNVs detectable in plasma are compared using violin plots in which horizontal dashed lines depict the median and interquartile range. All mutations identified using the population-based lung cancer-focused CAPP-Seq panel are considered. (**f**) The fraction of all mutant and wild-type cfDNA molecules (defined as in Fig. 1d) with fragment sizes falling within the size windows found to be ctDNA-enriched in Fig. 1e. (**g**) Violin plot displaying the enrichment of SNV VAFs following *in silico* size selection for the cfDNA fragment sizes found to be ctDNA-enriched in Fig. 1e. Enrichment is defined as the ratio of the SNV VAF following size selection to that observed prior to size selection. All mutations detectable in plasma prior to size selection (n=323 mutations) were considered. In the boxplot the center line denotes the median, the box contains the interquartile range, and the whiskers denote the extrema that are no more than 1.5 × IQR from the edge of the box (Tukey style). (**h**) Comparison of SNV VAFs before and after size selection. The dot plot displays the VAF of SNVs in plasma before and after size selection. The bar plot depicts the fraction of SNVs for which the VAF increased, decreased, or became un-detectable following size selection. All mutations detectable in plasma prior to size selection were considered. (**i**) Comparison of SNV VAFs prior to size selection in SNVs for which the VAF increased, decreased, or became un-detectable following size selection. (**j**) Tumor-informed ctDNA detection rates before and after size selection in patients sequenced with the population-based lung cancer-focused CAPP-Seq panel (n=85 patients) and customized capture panels designed using tumor exome sequencing data (n=17 patients).



Extended Data Figure 6. Clinical correlates of tumor-informed ctDNA detection.

(**a**) Relationship between metabolic tumor volume (MTV) measured by PET-CT and pretreatment ctDNA concentration measured in haploid genome equivalents per mL plasma (hGE/mL). All patients with detectable ctDNA and MTV measurements available were considered (n=46). Comparison performed by Spearman correlation. (**b**) Comparison of MTV in patients with and without detectable ctDNA. All patients with MTV measurements (n=81) were considered. (**c**) Multivariable linear regression was performed to associate the predictor variables (MTV, histology, and stage) with mean ctDNA VAF. For patients without detectable ctDNA, a VAF of 0.001% was used. All patients with MTV measurements (n=81) were considered. Additional details are provided in the Methods. (**d**) Comparison of pretreatment ctDNA levels in patients with adenocarcinoma histology and varying amounts of ground glass opacity (GGO) on pre-treatment CT scans. Brackets above depict comparison by Fisher’s Exact Test for ctDNA detection in patients with < 25% GGO (24/48 patients with ctDNA detected) vs. those with ≥ 25% GGO (2/13 patients with ctDNA detected). ND = not detected. All patients with adenocarcinoma histology and pre-treatment CT scans available were considered (n=61). (**e**) ctDNA detection rates in all patients (n=82, blue bars) and only those with adenocarcinoma histology (n=61, grey bars) with tumors that do or do not have evidence of necrosis on pre-treatment CT scans. Detection rates were compared by Fisher’s Exact Test. All patients with pre-treatment CT scans available were considered (n=82).



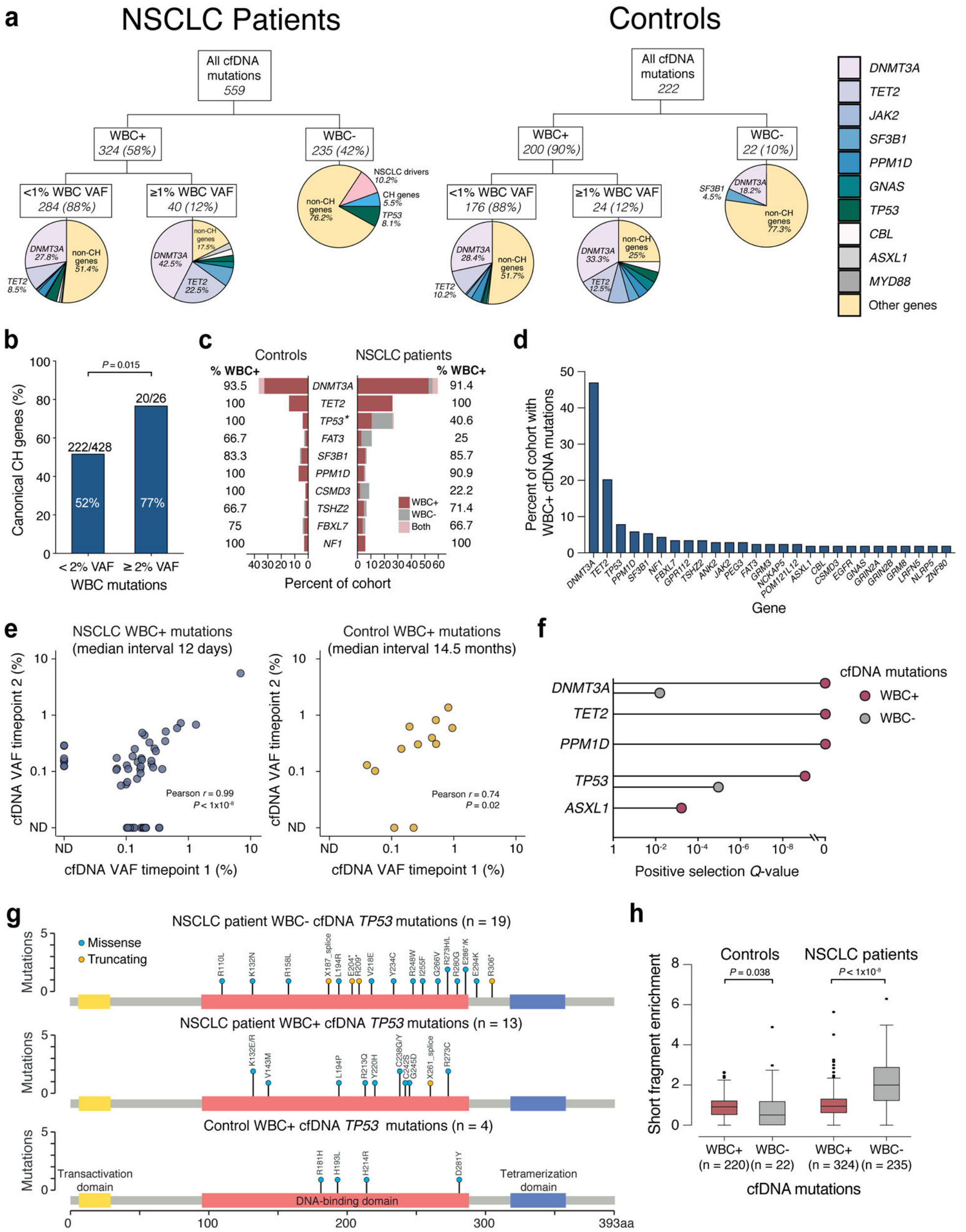
Parameter	Univariable		Multivariable	
	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value
Freedom from recurrence				
ctDNA - log10(VAF)	2.30 (1.50 to 3.54)	**0.0001**	2.52 (1.30 to 4.88)	**0.006**
Metabolic tumor volume - log10(mL)	1.88 (1.16 to 3.06)	**0.011**	1.41 (0.75 to 2.65)	0.287
Stage (I vs II-III)	0.59 (0.26 to 1.31)	0.195	2.50 (0.72 to 8.75)	0.151
Treatment (surgery only vs other)	0.45 (0.15 to 1.33)	0.148	0.70 (0.19 to 2.58)	0.594
Freedom from metastasis				
ctDNA - log10(VAF)	2.52 (1.51 to 4.21)	**0.0004**	2.76 (1.32 to 5.80)	**0.007**
Metabolic tumor volume - log10(mL)	1.91 (1.11 to 3.28)	**0.019**	1.28 (0.65 to 2.52)	0.471
Stage (I vs II-III)	0.58 (0.23 to 1.46)	0.245	1.86 (0.50 to 6.91)	0.357
Treatment (surgery only vs other)	0.45 (0.13 to 1.57)	0.209	0.96 (0.23 to 3.99)	0.952
Recurrence free survival				
ctDNA - log10(VAF)	2.15 (1.43 to 3.24)	**0.0002**	2.15 (1.17 to 3.96)	**0.014**
Metabolic tumor volume - log10(mL)	1.89 (1.19 to 3.01)	**0.0073**	1.50 (0.82 to 2.73)	0.186
Stage (I vs II-III)	0.58 (0.27 to 1.25)	0.164	2.06 (0.65 to 6.53)	0.220
Treatment (surgery only vs other)	0.49 (0.17 to 1.43)	0.192	0.81 (0.23 to 2.86)	0.745
Metastasis free survival				
ctDNA - log10(VAF)	2.26 (1.40 to 3.64)	**0.0008**	2.24 (1.15 to 4.37)	**0.018**
Metabolic tumor volume - log10(mL)	1.91 (1.14 to 3.18)	**0.0134**	1.40 (0.74 to 2.66)	0.303
Stage (I vs II-III)	0.57 (0.24 to 1.39)	0.216	1.53 (0.46 to 5.11)	0.489
Treatment (surgery only vs other)	0.51 (0.15 to 1.76)	0.286	1.10 (0.28 to 4.43)	0.889
Overall survival				
ctDNA - log10(VAF)	1.58 (0.99 to 2.51)	0.056	1.91 (0.87 to 4.19)	0.106
Metabolic tumor volume - log10(mL)	1.86 (1.06 to 3.24)	**0.029**	1.68 (0.75 to 3.79)	0.208
Stage (I vs II-III)	1.19 (0.40 to 3.51)	0.752	4.00 (0.74 to 21.6)	0.107
Treatment (surgery only vs other)	0.48 (0.11 to 2.18)	0.340	0.52 (0.09 to 3.14)	0.475

Cox proportional-hazards models

	High ctDNA		Low ctDNA	
Stage	IB	IB	IB	IA
Histology	Adenocarcinoma	Squamous	Adenocarcinoma	Adenocarcinoma
Tumor diameter	3.4 cm	3.5 cm	5.0 cm	2.9 cm
Pre-tx ctDNA	0.04%	0.014%	ctDNA not detected	ctDNA not detected
Follow up	Skeletal metastasis at 207 days	Peritoneal carcinomatosis at 547 days	Alive without disease at 1,461 days	Alive without disease at 1,160 days

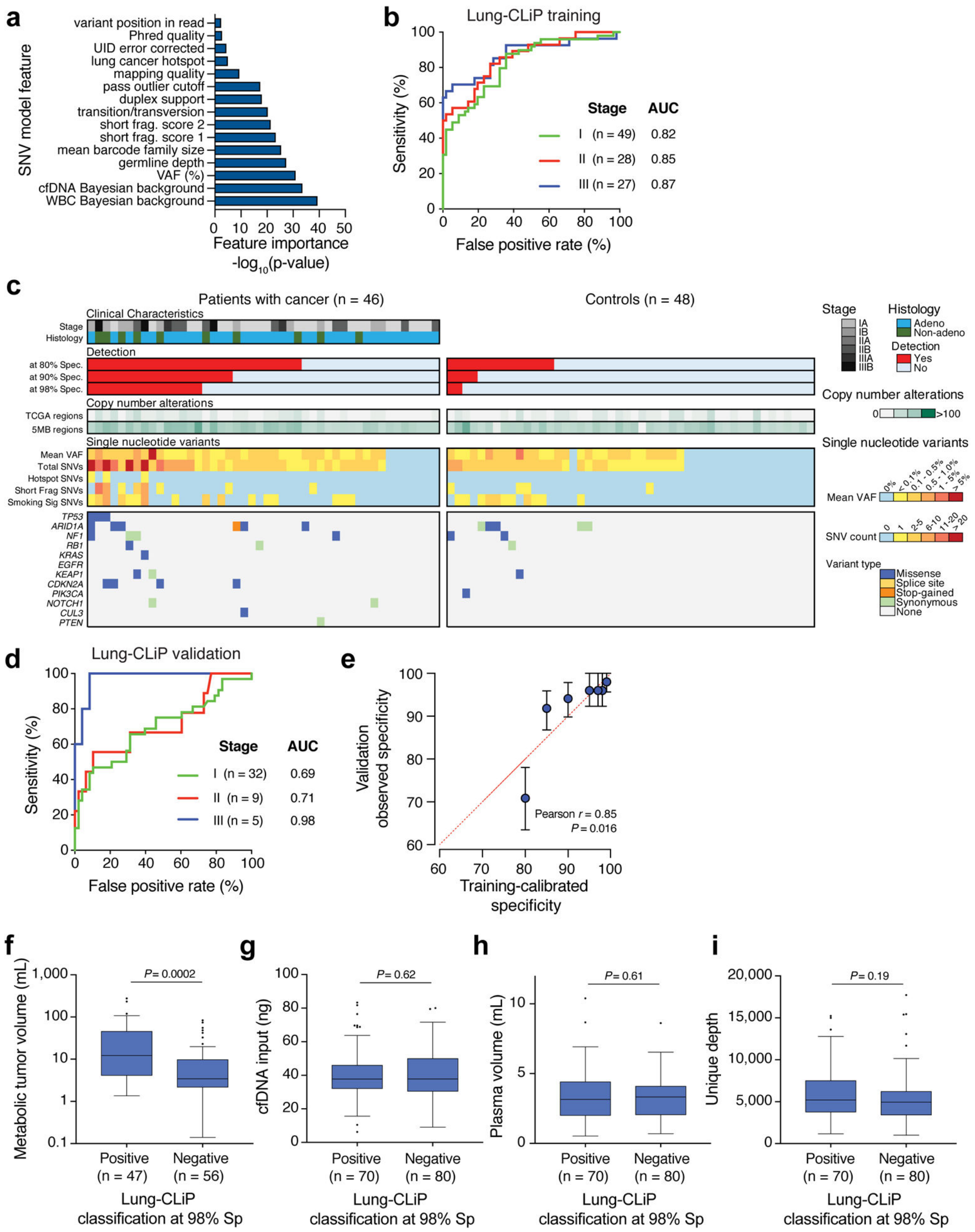
Extended Data Figure 7. Pretreatment ctDNA burden is prognostic in early-stage NSCLC.

(**a-d**) Kaplan–Meier analysis for recurrence-free survival (**a**,**b**) and freedom from metastasis (**c**,**d**) stratified by pretreatment ctDNA level in all stage I-III patients (**a,c**, n=85) and stage I patients only (**b,d,** n=48). The median ctDNA level across the cohort (0.0031%) was used to stratify patients into ctDNA high and ctDNA low groups. *P*-values were calculated using the log-rank test. HR = hazard ratio. (**e**) Table summarizing the results of univariable and multivariable Cox proportional hazards models. Metabolic tumor volume (MTV) measured by PET-CT and ctDNA measurements (mean SNV VAF) were log transformed. Significant *P*-values (< 0.05) are bolded. For univariable analysis of ctDNA level and stage, all patients (n=85) were considered. For the univariable analysis of MTV, and for all multivariable analysis, only patients with MTV measurements available (n=81) were considered. Univariable and multivariable *P*-values were assessed using the log-likelihood test. (**f**) Example patients with stage I adenocarcinoma. On the left are two patients with high pretreatment ctDNA levels who developed distant metastases following surgery. On the right are two patients with undetectable ctDNA who achieved long term remissions following surgery.



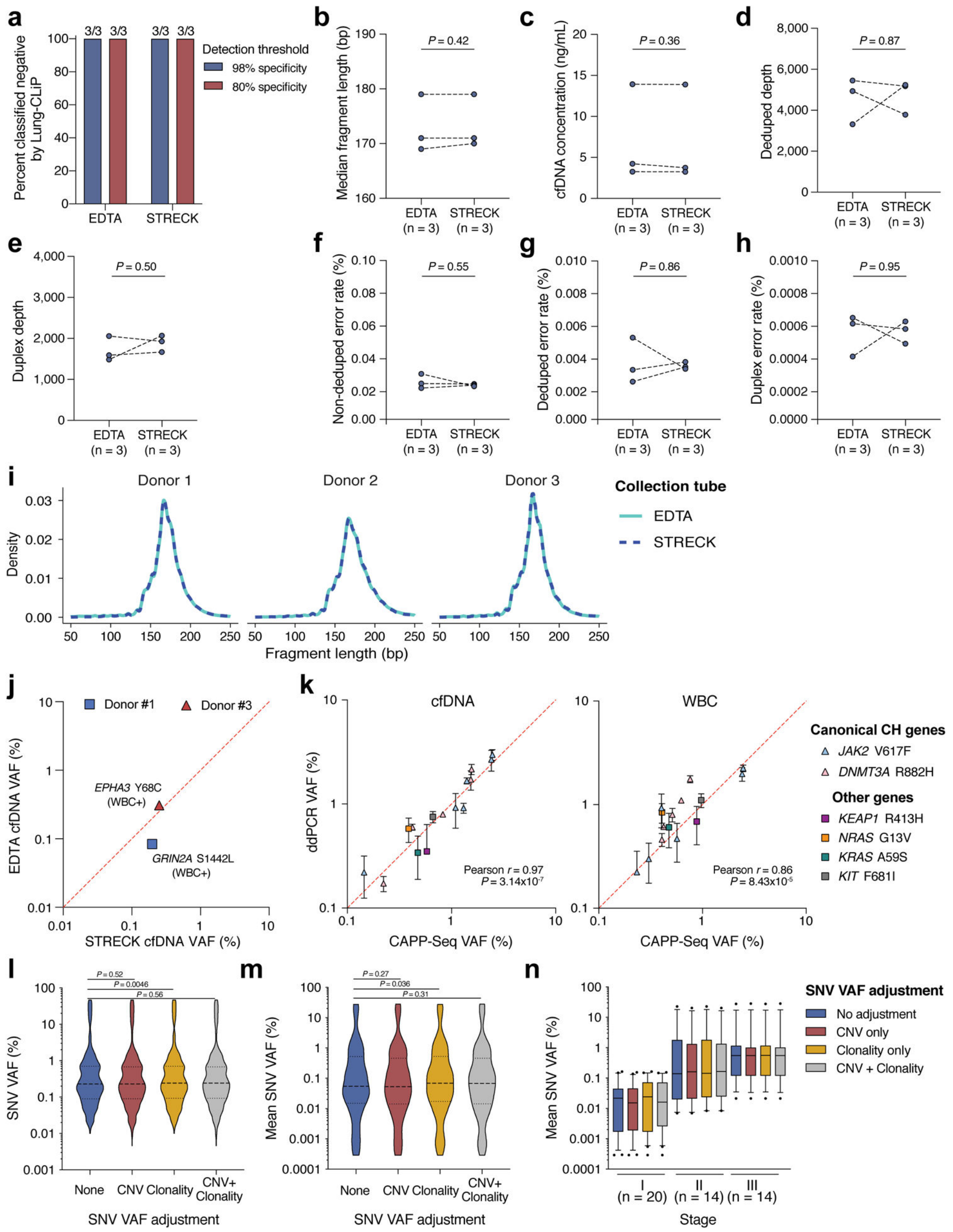
Extended Data Figure 8. Biological features of cfDNA mutations reflecting clonal hematopoiesis.

(**a**) Flow chart depicting the fraction of WBC+ and WBC- cfDNA mutations affecting canonical CH genes in NSCLC patients and controls. WBC+ cfDNA mutations present at ≥ 1% VAF in matched leukocytes more frequently affect canonical CH genes than those below 1% (51/64 vs. 223/460 WBC+ cfDNA mutations present at ≥ 1% vs. < 1% VAF in matched leukocytes affect canonical CH genes, respectively; $P = 1.9\times10^{-6}$ Fisher's Exact Test). Only mutations identified *de novo* in the cfDNA for which presence in the matched WBCs could be confidently assessed are considered (Methods). (**b**) The percent of mutations genotyped *de novo* from WBC DNA at VAFs < 2% and ≥ 2% affecting canonical CH genes in patients and controls (all patients and controls are considered. Comparison was performed by Fisher's Exact Test. (**c**) The percent of controls (left) and patients (right) with one or more mutations in the 10 genes that most frequently contained WBC+ cfDNA mutations. NSCLC patients and controls with only WBC+ mutations, only WBC- mutations, or both WBC+ and WBC- mutations in a gene are depicted in red, grey, and pink, respectively. The numbers next to each bar represent the percent of all cfDNA mutations in that gene that are WBC+ in NSCLC patients (right) or controls (left). NSCLC patients had significantly more WBC- cfDNA mutations in *TP53* than controls (19/32 vs. 0/4 in patients vs. controls, respectively. * = Fisher's Exact Test, $P = 0.04$). (**d**) Mutation frequency by gene for WBC+ cfDNA mutations observed across all NSCLC patients (n=104) and controls (n=98). The y-axis depicts the percent of the combined cohort with WBC+ cfDNA mutations affecting a given gene. All genes with mutations in 4 or more individuals in the combined cohort are depicted. (**e**) Scatterplot comparing the VAFs of WBC+ cfDNA mutations across multiple timepoints in NSCLC patients (left panel, n=54 mutations, n=8 individuals) and controls (right panel, n=12 mutations, n=6 individuals). Statistical comparison was performed by Pearson correlation on mutations detected at both time points. (**f**) Positive selection analysis was carried out on all synonymous and nonsynonymous WBC+ (n=693 mutations, red) and WBC- (n=526 mutations, grey) cfDNA mutations observed in NSCLC patients and controls using the dNdScv R package with a modification to account for the fraction of a given gene covered by our sequencing panel. The x-axis indicates the dNdScv adjusted *P*-value (Q-value) for all substitution types. Genes were considered under positive selection if the Q-value was < 0.05. All genes meeting this threshold are displayed. Additional details are provided in the Methods. (**g**) distribution of WBC+ and WBC- cfDNA mutations across the p53 protein in NSCLC patients and controls. (**h**) Short fragment enrichment of WBC+ and WBC- cfDNA mutations in NSCLC patients and controls, defined as the fold change in VAF for a given mutation following *in silico* size selection for the cfDNA fragment sizes found to be ctDNA-enriched in Fig. 1e. The center line denotes the median, the box contains the interquartile range, and the whiskers denote the 10th and 90th percentile values.



Extended Data Figure 9. Feature importance and performance of Lung-CLiP.

(**a**) Biological and technical parameters specific to each individual variant used as features in a dedicated logistic regression ‘SNV model’. The feature names are depicted on the y-axis and the negative log10 of the *P*-value derived from comparing all post-filtered SNVs in NSCLC patients (n=574 mutations from n=104 individuals) vs. those in risk-matched controls (n=64 mutations from n=56 individuals) in a univariable linear model in the training set is shown on the x-axis. All features with a *P*-value < 0.01 are shown, *P*-values were calculated using an un-paired two-sided t-test. Additional information about each feature is provided in the Supplemental Methods. (**b**) Receiver operator characteristic (ROC) curves for the Lung-CLiP model depicting performance stratified by tumor stage in the training set (n=104 NSCLC patients and n=56 risk-matched controls). (**c**) Spectrum of clinicopathologic correlates and selected features observed across the 46 early-stage NSCLC patients and 48 risk-matched controls undergoing annual lung cancer screening in a prospectively enrolled independent validation cohort. (**d**) ROC curves for the Lung-CLiP model depicting performance stratified by tumor stage in the validation set (n=46 NSCLC patients and n=48 risk-matched controls). (**e**) Comparison of the specificity observed in the validation cohort at different thresholds defined in the training cohort. Dots denote the median specificity across 1,000 bootstrap re-samplings and error bars depict the interquartile range. Statistical comparison was performed by Pearson correlation on the non-bootstrapped data. (**f-i**) Comparison of (**f**) metabolic tumor volume, (**g**) cfDNA input to library preparation, (**h**) plasma volume used, and (**i**) unique sequencing depth in NSCLC patients correctly classified at 98% specificity (“Positive”) to those in patients incorrectly classified (“Negative”). All NSCLC patients in the training and validation cohorts were considered (n=103 patients with metabolic tumor volume measurements in **f** and n=150 patients in **g-i** and). In box plots the center line denotes the median, the box contains the interquartile range, and the whiskers denote the extrema that are no more than 1.5 × IQR from the edge of the box (Tukey style).



Extended Data Figure 10. Technical reproducibility and benchmarking of CAPP-Seq and the Lung-CLiP model.

(a-j) Blood was drawn from each of three healthy donors into two STRECK tubes and two K_2EDTA tubes and processed using the protocols used in our study. cfDNA extraction and library preparation were performed as described in the Methods with 25 ng of cfDNA input for each sample. Sequencing and data processing were performed as described in the Methods and each sample was downsampled to 80 million reads prior to barcode-deduplication to facilitate comparison. (**a**) The Lung-CLiP model was trained on the 104 NSCLC patients and 56 risk-matched controls in the training cohort and applied to the cfDNA samples extracted from plasma drawn into STRECK and K_2EDTA tubes. The fraction of donors classified as negative by Lung-CLiP at the 98% (blue bars) and 80% (red bars) specificity thresholds defined in the training data are depicted. (**b**-**h**) Comparison of (**b**) median cfDNA fragment size, (**c**) cfDNA concentration in ng/ml, (**d**) deduped depth, (**e**) duplex depth, and (**f-h**) error metrics in cfDNA samples extracted from plasma drawn into the two tube types. cfDNA samples from the same donor are connected with dashed lines, comparisons were performed using a paired two-sided t-test. (**i**) Comparison of the fragment size distribution of cfDNA samples extracted into the two tube types. (**j**) Genotyping was performed as described in the Methods on cfDNA samples extracted from plasma drawn into the two tube types from the three donors. Donor #1 and donor #3 each had one mutation identified in cfDNA which was present in samples extracted from plasma drawn into both tube types and was also present in matched WBCs (WBC+). Donor #2 had no mutations identified in cfDNA samples extracted from plasma drawn into either tube type. (**k**) Orthogonal validation of WBC+ cfDNA mutations (n=15) using droplet digital PCR (ddPCR). Comparison of the VAF of WBC+ cfDNA mutations as measured by CAPP-Seq (x-axis) and ddPCR (y-axis). ddPCR was performed in triplicate on cfDNA (left) or WBC DNA (right) sequencing libraries. All 15 mutations (100%) were validated by ddPCR in both the cfDNA and WBC compartments. Triangles represent recurrent “hotspot” mutations in canonical CH genes and squares represent private mutations in non-CH genes. Statistical comparison was performed by Pearson correlation. (**l-n**) Tumor-informed ctDNA levels in NSCLC patients with and without adjustments for copy number state and clonality of tumor mutations. (**l**) VAFs of individual mutations (n=323) observed in cfDNA with different SNV VAF adjustment strategies. Comparisons were performed using a paired two-sided t-test. (**m**) The mean cfDNA VAF across all tracked mutations tracked in patients with detectable ctDNA (n=48) with the different adjustment strategies. Comparisons were performed using a paired two-sided t-test. (**n**) The same data as in **m** separated by stage. In box plots the center line denotes the median, the box contains the interquartile range, and the whiskers denote the extrema that are no more than 1.5 × IQR from the edge of the box (Tukey style). In **l-n**, copy number and clonality adjustment was performed as described in the Supplementary Methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Eric Kool (PhD) for his advice relating to ROS scavengers and Eric Edell and Aaron Bungum from the Mayo Clinic Lung Tumor Specimen Registry for their assistance with sample collection. This work was supported by the National Cancer Institute (R01CA188298 and R01CA233975 to M.D. and A.A.A., 1-K08-CA241076-01 to D.M.K., R25CA180993 and T32-CA 121940 to M.S.E., U01 CA196405 to P.P.M., training grant T32 CA009302 to E.G.H., M.C.N and D.A., and K12CA090628 and P30 CA015083-44S1 to A.S.M.), the US National Institutes of Health Director’s New Innovator Award Program (1-DP2-CA186569, M.D.), the US National Institutes of Health, the Virginia and D.K. Ludwig Fund for Cancer Research (M.D. and A.A.A), the CRK Faculty Scholar Fund (M.D.), the Bakewell Foundation (M.D. and A.A.A.), the Damon Runyon Cancer Research Foundation (PST#09-16, D.M.K.), the Tobacco-Related Disease Research Program Predoctoral Fellowship (T30DT0806, E.G.H) the Blavatnik Family Fellowship (E.G.H), the American Cancer Society (134031-PF-19-164-01-TBG, B.Y.N.), the SDW/DT and Shanahan Family Foundations (A.A.A.), Stand Up To Cancer (M.D., A.A.A, D.A.H. and L.V.S.), and the NSF Graduate Research Fellowship (DGE-114747 to J.J.C, DGE-1656518 to D.A.). A.A.A. is a Scholar of The Leukemia & Lymphoma Society.

Competing Interests Statement

D.M.K reports paid consultancy from Roche Molecular Diagnostics. A.A.C. reports speaker honoraria and travel support from Roche Sequencing Solutions, Varian Medical Systems, and Foundation Medicine and a research grant from Roche Sequencing Solutions and has served as a paid consultant for Fenix Group International. A.S.M. reports advisory for AbbVie, Genentech, and Bristol-Myers Squibb (honoraria paid to institution) and research funding from Novartis and Verily. J.J. is now employed by Celgene. S.L.H. reports paid advisory from AstraZeneca, speaker honoraria from Varian Medical System and research funding from BeyondSpring Pharmaceuticals Inc., Hitachi Chemical Diagnostics, Genentech, and New River Labs. S.S.G reports paid consultancy from AbbVie, Ceremark, CytomX Therapeutics Inc, GPV, Life Molecular Imaging, Nusano, Spectrum Dynamics and TPG and ownership interest in Akrotome Imaging Inc, Cellsight Technologies, CytomX Therapeutics, Inc, Earli Inc, Endra Inc, MagArray Inc, Nines, Nodus Therapeutics, Nusano, RefleXion Medical Inc, SiteOne Therapeutics Inc, Spectrum Dynamics, Vave Health and Vor Biopharma. J.W.N reports paid consultancy from AstraZeneca, Genentech, Roche, Exelixis, Jounce Therapeutics, Takeda Pharmaceuticals, Eli Lilly and Company, Calithera Biosciences and research funding from Genentech, Roche, Merck, Novartis, Boehringer Ingelheim, Exelixis, Nektar Therapeutics, Takeda Pharmaceuticals, Adaptimmune and GSK. H.A.W. reports paid advisory from AstraZeneca, Xcovery, Janssen, and Mirati, un-paid advisory from Merck, Takeda, Genentech, Roche, Cellworks and research funding from ACEA Biosciences, Arrys Therapeutics, AstraZeneca/Medimmune, BMS, Celgene, Clovis Oncology, Exelixis, Genentech/Roche, Gilead, Lilly, Merck, Novartis, Pfizer, Pharmacyclics, and Xcovery. A.A.A. reports ownership interest in CiberMed and FortySeven Inc, patent filings related to cancer biomarkers, and paid consultancy from Genentech, Roche, Chugai, Gilead, and Celgene. M.D. reports research funding from Varian Medical Systems, ownership interest in CiberMed, patent filings related to cancer biomarkers, and paid consultancy from Roche, AstraZeneca, and BioNTech. The remaining authors declare no potential conflicts of interest.

References

1. National Lung Screening Trial Research Team et al. Results of initial low-dose computed tomographic screening for lung cancer. N. Engl. J. Med 368, 1980–91 (2013). [PubMed: 23697514]
2. de Koning HJ et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. N. Engl. J. Med NEJMoa1911793 (2020). doi:10.1056/NEJMoa1911793
3. Moyer VA Screening for Lung Cancer: U.S. Preventive Services Task Force Recommendation Statement. Ann. Intern. Med 160, 330–338 (2014). [PubMed: 24378917]
4. Jemal A & Fedewa SA Lung Cancer Screening With Low-Dose Computed Tomography in the United States—2010 to 2015. JAMA Oncol. 3, 1278 (2017). [PubMed: 28152136]
5. Doria-Rose VP et al. Use of lung cancer screening tests in the United States: Results from the 2010 National Health Interview Survey. Cancer Epidemiol. Biomarkers Prev 21, 1049–1059 (2012). [PubMed: 22573798]
6. Newman AM et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat. Biotechnol 34, 547–555 (2016). [PubMed: 27018799]
7. Chaudhuri AA et al. Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. Cancer Discov. 7, 1394–1403 (2017). [PubMed: 28899864]
8. Abbosh C et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature 545, 446–451 (2017). [PubMed: 28445469]
9. Jiang P et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc. Natl. Acad. Sci 112, E1317–E1325 (2015). [PubMed: 25646427]
10. Mouliere F et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci. Transl. Med 10, eaat4921 (2018). [PubMed: 30404863]
11. Travis WD et al. International association for the study of lung cancer/American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma. J. Thorac. Oncol 6, 244–285 (2011). [PubMed: 21252716]
12. Moding EJ et al. Circulating tumor DNA dynamics predict benefit from consolidation immunotherapy in locally advanced non-small-cell lung cancer. Nat. Cancer (2020). doi:10.1038/s43018-019-0011-0
13. Steensma DP et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood 126, 9–16 (2015). [PubMed: 25931582]
14. Lui YYN et al. Predominant hematopoietic origin of cell-free dna in plasma and serum after sex-mismatched bone marrow transplantation. Clin. Chem 48, 421–427 (2002). [PubMed: 11861434]

15. Liu J et al. Biological background of the genomic variations of cf-DNA in healthy individuals. Ann. Oncol 1–7 (2018). doi:10.1093/annonc/mdy513

16. Razavi P et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. Nat. Med 25, (2019).

17. Ptashkin RN et al. Prevalence of Clonal Hematopoiesis Mutations in Tumor-Only Clinical Genomic Profiling of Solid Tumors. JAMA Oncol. 4, 1589 (2018). [PubMed: 29872864]

18. Hammerman PS et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489, 519–525 (2012). [PubMed: 22960745]

19. Collisson EA et al. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. Nature 511, 543–550 (2014). [PubMed: 25079552]

20. Alexandrov LB et al. Signatures of mutational processes in human cancer. Nature 500, 415–21 (2013). [PubMed: 23945592]

21. Hainaut P & Pfeifer GP Somatic TP53 mutations in the era of genome sequencing. Cold Spring Harb. Perspect. Med 6, (2016).

22. Shen SY et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 563, 579–583 (2018). [PubMed: 30429608]

23. Cohen JD et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 359, 926–930 (2018). [PubMed: 29348365]

24. Phallen J et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci. Transl. Med 9, eaan2415 (2017). [PubMed: 28814544]

25. Cristiano S et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389 (2019). [PubMed: 31142840]

26. Simon R Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers. J. Clin. Oncol 23, 7332–7341 (2005). [PubMed: 16145063]

27. Ma J, Ward EM, Smith R & Jemal A Annual number of lung cancer deaths potentially avertable by screening in the United States. Cancer 119, 1381–1385 (2013). [PubMed: 23440730]

28. Kurtz DM et al. Dynamic Risk Profiling Using Serial Tumor Biomarkers for Personalized Outcome Prediction. Cell 1–15 (2019). doi:10.1016/j.cell.2019.06.011

29. Chen S et al. AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics 18, (2017).

30. Li H & Durbin R Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009). [PubMed: 19451168]

31. Karczewski KJ et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210 (2019). doi:10.1101/531210

32. Koboldt DC et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576 (2012). [PubMed: 22300766]

33. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol 31, 213–219 (2013). [PubMed: 23396013]

34. Saunders CT et al. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012). [PubMed: 22581179]

35. Carter SL et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol 30, 413–421 (2012). [PubMed: 22544022]

36. Bailey MH et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 173, 371–385.e18 (2018). [PubMed: 29625053]

37. Martincorena I et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171, 1029–1041.e21 (2017). [PubMed: 29056346]

38. Rosenthal R, McGranahan N, Herrero J, Taylor BS & Swanton C deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 17, 1–11 (2016). [PubMed: 26753840]

39. Hindson BJ et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal. Chem 83, 8604–8610 (2011). [PubMed: 22035192]

40. Mermel CH et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, 1–14 (2011).

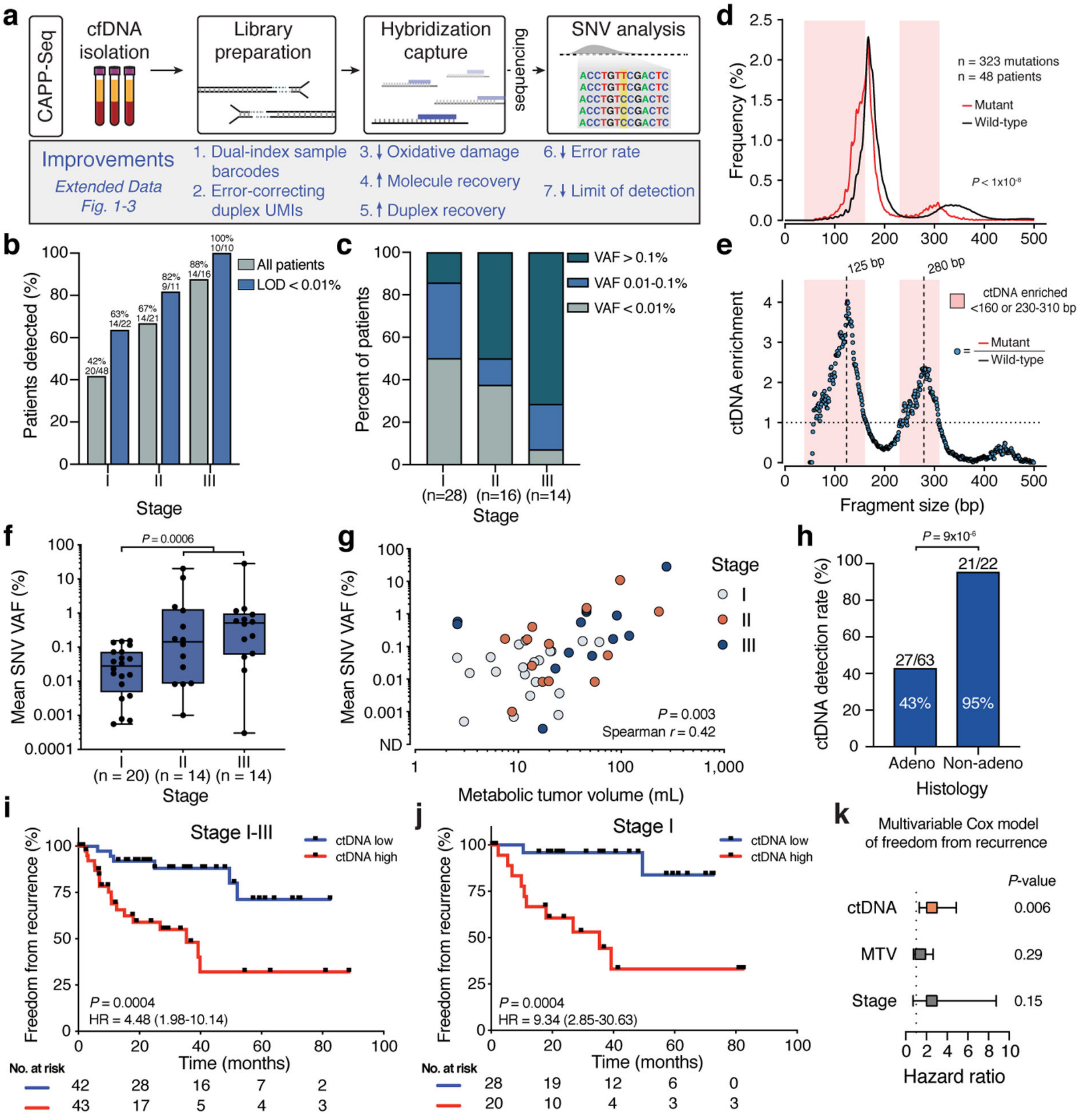


Figure 1. Biological and clinical correlates of ctDNA burden in early-stage lung cancer patients.

(a) Summary of key methodical improvements to the CAPP-Seq workflow. (b) Tumor-informed ctDNA detection rates across all patients (grey bars, n=85) and the subset of patients with an analytical limit of detection (LOD) < 0.01% (blue bars, n=43). (c) Pretreatment ctDNA levels, quantified as the mean variant allele frequency (VAF) across all mutations tracked, summarized by stage in NSCLC patients with detectable ctDNA or a LOD < 0.01%. (d) Fragment size distribution of cfDNA molecules containing mutations present in matched tumor samples (red line) and wild-type molecules overlapping the same genomic positions in the same patients (black line). Size distributions were compared by the Kolmogorov-Smirnov test. Fragment size regions enriched for ctDNA are shaded in red. (e) The relative enrichment of mutant vs. wild-type cfDNA molecules (i.e. "ctDNA enrichment") calculated from the data depicted in panel d. Fragment size regions enriched for ctDNA are shaded in red. (f) Pretreatment ctDNA levels summarized by stage in patients with detectable ctDNA. Brackets depict comparison of stage I (n=20) vs. stage II-III (n=28) patients. (g) Relationship between metabolic tumor volume (MTV) and pretreatment ctDNA level. All patients with detectable ctDNA and MTV measurements available were considered (n=46). Comparison performed by Spearman correlation. (h) ctDNA detection rates in patients with adenocarcinoma and non-adenocarcinoma histology. Comparison performed by Fisher's Exact Test. (i-j) Kaplan–Meier analysis for freedom from recurrence stratified by pretreatment ctDNA level in (i) all stage I-III patients (n=85) and (j) stage I patients only (n=48). The median ctDNA level across the cohort (0.0031%) was used to stratify patients into ctDNA low and ctDNA high groups. HR=hazard ratio. (k) Results of multivariable Cox proportional hazards model for freedom from recurrence in patients with MTV measurements available (n=81). Points denote the hazard ratio and error bars depict the 95% CI.

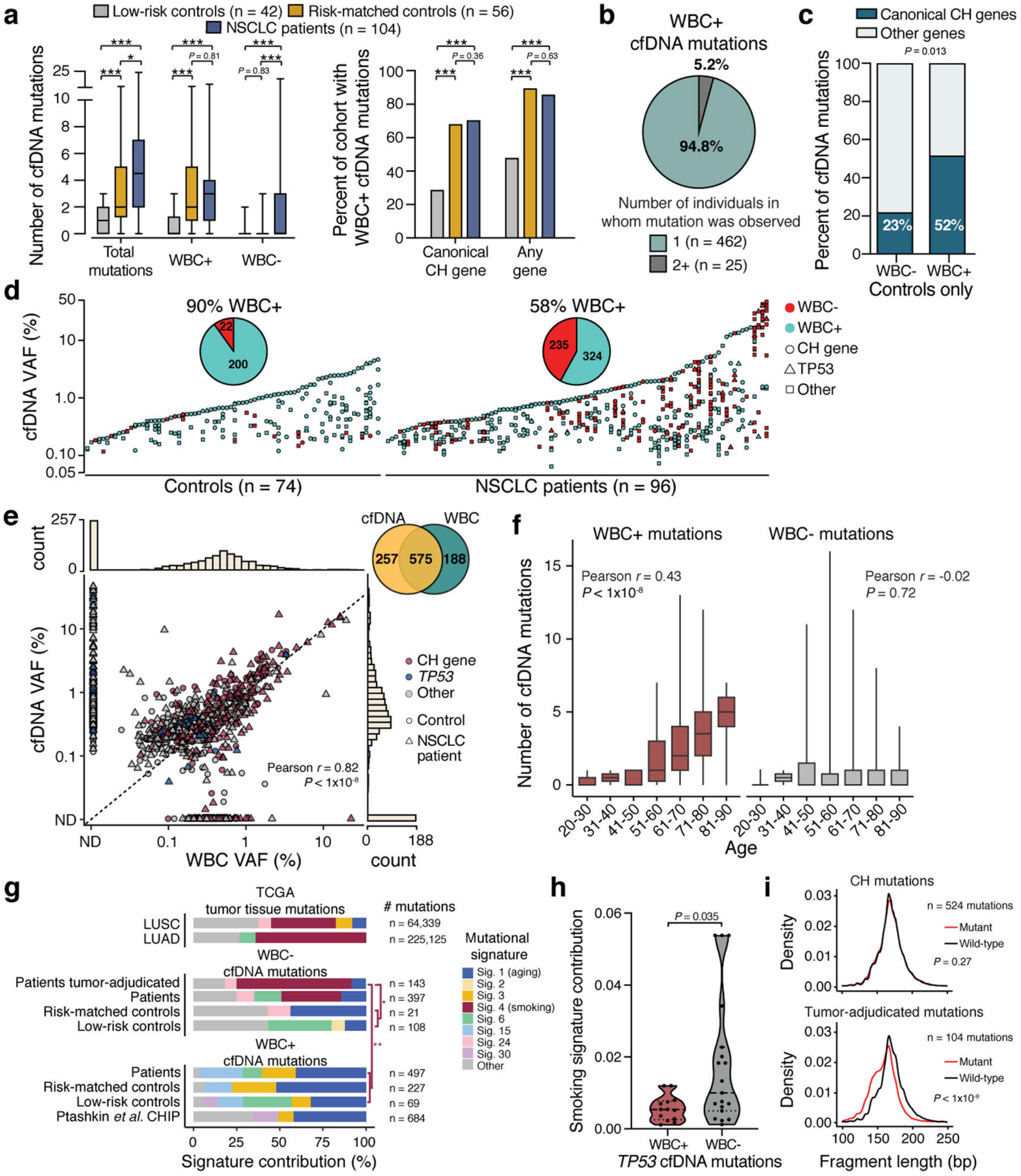


Figure 2. Clonal hematopoiesis (CH) is a major source of cfDNA variants and molecular features distinguish CH-derived from tumor-derived cfDNA variants.

(**a**) *Left*: Count of total, WBC+, and WBC- nonsynonymous cfDNA mutations in NSCLC patients, risk-matched controls, and low-risk controls. (*, $P < 0.01$; **, $P < 0.001$; ***, $P < 0.0001$). *Right*: Percent of each cohort with one or more WBC+ cfDNA mutations in a canonical CH gene or in any gene. Comparisons performed using Fisher's Exact Test (***, $P < 1\times10^{-5}$). (**b**) Percent of WBC+ cfDNA mutations that were private vs. those observed in two or more individuals. All NSCLC patients and controls were considered (n=202). (**c**) Percent of WBC- and WBC+ cfDNA mutations affecting canonical CH genes vs. other genes in controls. Comparison performed using Fisher's Exact Test between WBC+ (n=200) and WBC- (n=22) cfDNA mutations. All controls were considered (n=98). (**d**) Variant allele frequencies (VAFs) of cfDNA mutations observed in controls (left), and NSCLC patients (right). The color denotes whether a cfDNA mutation was WBC- (red) or WBC+ (blue) and the shape denotes the type of gene. Individuals with one or more cfDNA mutations are shown. Pie charts display counts of WBC- and WBC+ cfDNA mutations pooled by cohort. (**e**) Scatterplot depicting the VAFs of mutations in cfDNA and matched WBCs. The color denotes the type of gene and the shape denotes whether the mutation was observed in a NSCLC patient or control. All mutations genotyped *de novo* in the cfDNA or WBCs for which presence in the other compartment could be confidently assessed are shown. Marginal histograms display the VAF distribution of all mutations in cfDNA or WBCs. Comparison performed by Pearson correlation on mutations detected in both compartments (n=575). (**f**) Association between age and number of WBC+ or WBC- cfDNA mutations. All patients (n=104) and controls (n=98) were considered. Comparison performed by Pearson correlation on the un-binned data. (**g**) Mutational signature contributions in WBC+ and WBC- cfDNA mutations in NSCLC patients and controls compared to the CH and lung cancer literature[17–19]. Statistical significance was assessed for differences in signature 4 (smoking) as described in the Methods (*, $P = 0.005$; **, $P < 1\times10^{-8}$). (**h**) Smoking signature contribution in WBC+ (n=13) vs. WBC- (n=19) *TP53* mutations in NSCLC patients. (**i**) Fragment size distributions of cfDNA molecules containing mutations present in matched WBC DNA (“CH mutations,” top) or matched tumor samples (“tumor-adjudicated,” bottom) compared to wild-type cfDNA molecules overlapping the same genomic positions in the same patients. Size distributions compared using the Kolmogorov-Smirnov test.

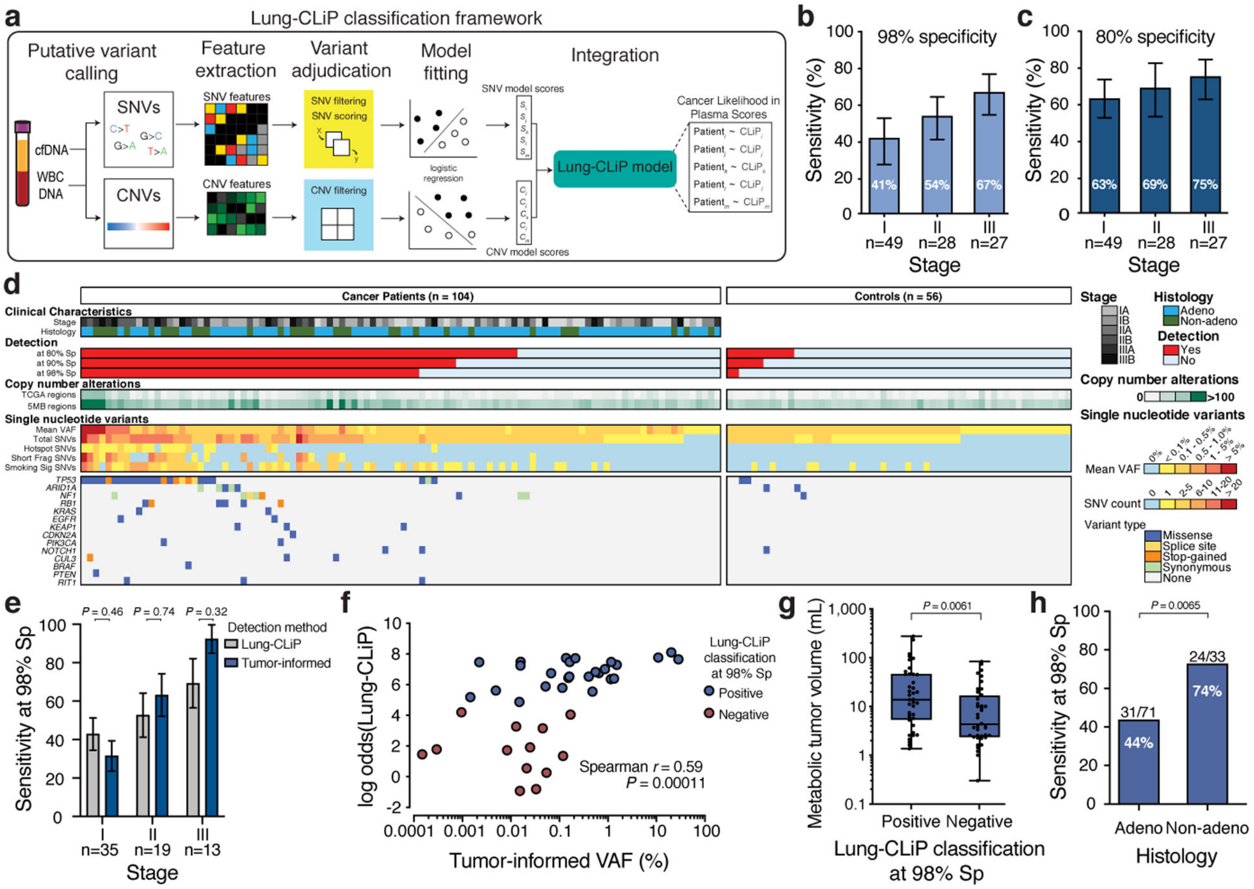


Figure 3. Development of the Lung Cancer Likelihood in Plasma (Lung-CLiP) method.

(**a**) Schematic of the Lung-CLiP classification framework. (**b-c**) Sensitivity of detection by stage at (**b**) 98% and (**c**) 80% specificity as determined in a leave-one-out cross validation in the training cohort. Bars denote the median sensitivity across 1,000 bootstrap re-samplings and error bars depict the interquartile range. (**d**) Clinicopathologic correlates and selected molecular features observed in the NSCLC patients and risk-matched controls undergoing annual lung cancer screening in the training cohort. (**e**) Sensitivity of ctDNA detection summarized by stage using tumor-informed CAPP-Seq and Lung-CLiP in patients with matched tumor tissue (n=67). Detection thresholds achieving ≥ 98% specificity were used for both approaches. Data is depicted as in panels **b-c**. Sensitivity comparisons performed by Fisher's Exact Test on the non-bootstrapped data. (**f**) Relationship between ctDNA level and Lung-CLiP score in patients with detectable ctDNA by tumor-informed CAPP-Seq (n=39). The x-axis depicts the mean variant allele frequency (VAF) across all mutations tracked by tumor-informed CAPP-Seq and the y-axis depicts the log odds of the Lung-CLiP score. Comparison performed by Spearman correlation. (**g**) Metabolic tumor volume in NSCLC patients correctly classified at 98% specificity ("Positive," n=40) and those incorrectly classified ("Negative," n=40). (**h**) Sensitivity of detection by Lung-CLiP at 98% specificity in patients with adenocarcinoma vs. non-adenocarcinoma histology. Comparison performed by Fisher's Exact Test.

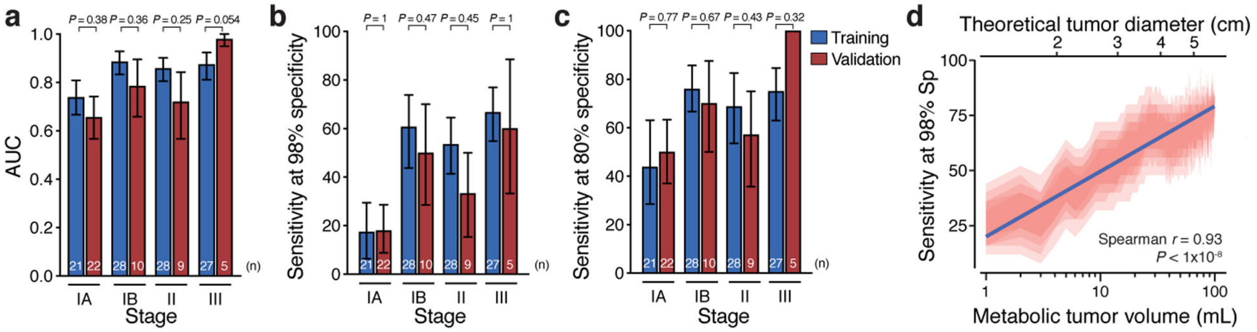


Figure 4. Validation of Lung-CLiP in a prospectively collected independent cohort.

(**a-c**) Comparison of (**a**) AUC and sensitivity at (**b**) 98% and (**c**) 80% specificity stratified by stage in the training (blue) and validation (red) cohorts. Bars denote the median value observed across 1,000 bootstrap re-samplings and error bars depict the interquartile range. AUC comparisons were performed using Delong's method and sensitivity comparisons were performed using Fisher's Exact Test on the non-bootstrapped data. The 98% and 80% specificity thresholds were defined in the training data. (**d**) Relationship between metabolic tumor volume (MTV) and sensitivity of Lung-CLiP at 98% specificity. Using 1,000 bootstrap re-samplings, sensitivity was calculated over a 25-patient sliding window of MTVs (lower x-axis). The upper x-axis depicts the theoretical tumor diameter of a single lesion corresponding to the MTVs on the lower x-axis assuming a spherical geometry. All NSCLC patients with MTV measurements in the training (n=80) and validation (n=23) were considered. The blue line represents a linear fit of $\log_{10}$(MTV) vs. sensitivity and red shaded regions depict the 95%, 85%, 75%, 65%, and 55% confidence intervals. Comparison of sensitivity in a given window to the average MTV in that window was performed by Spearman correlation using the non-bootstrapped data.